BITTIGER

DS 501 Data scientist express bootcamp

Week 2 [Ella]

BITTIGER

版权声明

BITTIGER

Copyright Policy

All content included on the Site or third-party platforms as part of the class, such as text, graphics, logos, button icons, images, audio clips, video clips, live streams, digital downloads, data compilations, and software, is the property of BitTiger or its content suppliers and protected by copyright laws.

Any attempt to redistribute or resell BitTiger content will result in the appropriate legal action being taken.

We thank you in advance for respecting our copyrighted content.

For more info:
see https://www.bittiger.io/termsofuse
and https://www.bittiger.io/termsofservice

Summary

- Hypothesis testing (cont'd)
  - Test for variance, Chi square test
- Simple linear regression
  - Modeling E(Y|X)
  - Assumption
  - Coefficient estimation
    - Least square estimation
    - Maximum likelihood estimation
    - Hypothesis testing

Summary

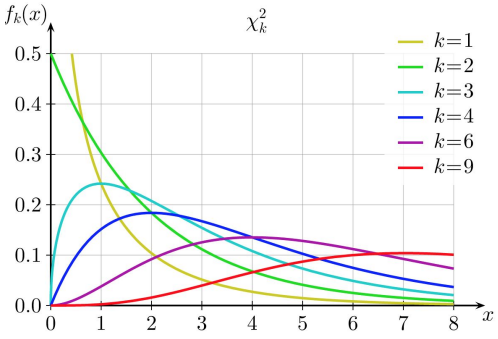
- Multiple linear regression
  - Coefficient estimation
  - Evaluate model performance
  - ANOVA and F test
- Residual
  - Residual diagnostics
  - Leverage, standardizing

Chi square distribution

- Definition
  - Sum of square of k standard normal random variables, N (0, 1)

$$\chi^2(k) = \sum_{i=1}^{k} Z_i^2$$

  - k-1 degree of freedom



- Chi square test for one variance

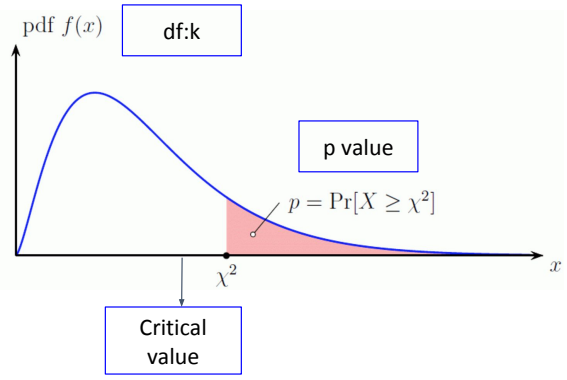
$H_0 : \sigma^2 = \sigma_0^2$ $\quad H_A : \sigma^2 \neq \sigma_0^2, \quad H_A : \sigma^2 < \sigma_0^2,$ or $H_A : \sigma^2 > \sigma_0^2$

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

http://uregina.ca/~gingrich/ch10.pdf
http://stats.stackexchange.com/questions/16921/how-to-understand-degrees-of-freedom/17148#17148
(use chi square context to explain df in test)

Decide to reject or not reject null hypothesis

- Compare p value with type I error
  - If p value is smaller, reject null hypothesis
- Compare chi square statistic with critical value
  - If chi square statistic is larger than critical value, reject null hypothesis
- Chi square test is always one sided test



What does it mean if we use chi square test as two sided test, meaning we would also be worried if the statistic were too far into the *left* side of the chi-squared distribution. This would mean that we are worried the fit might be *too good*.

What can Chi-square test be used for?

- One sample test
  - A population variance
  - Compare categorical variable with known distribution (goodness of fit)
  - Null and alternative hypothesis?
- Two sample test
  - Compare two population variance
  - Compare two categorical variables to test if they have same distribution.
  - Chi-square independence test
  - Null and alternative hypothesis?

Test about variance of two populations: https://onlinecourses.science.psu.edu/stat414/node/273

One sample
The variable follows expected distribution

Two sample
Ho: The two categorical variables are independent.
Ha: The two categorical variables are related.
http://ccnmtl.columbia.edu/projects/qmss/the_chisquare_test/about_the_chisquare_test.html
http://www2.lv.psu.edu/jxm57/irp/chisquar.html
http://practicalsurveys.com/reporting/chisquare.php

Chi square test for goodness of fit

- Test if X follows certain distribution F, X is categorical var
  - H_0: X follows F
    H_a: X does not follow F
  - Calculate (chi-square) statistics
    - Recall t/z score, which measures the (normalized) distance between observed stats and expected stats given H_0 is true
    - Similarly, $chi\ square = \sum_{all\ cells} \frac{(observed - expected)^2}{expected}$
    - Flip coin example

	head	tail
observed	0	10
expected	5	5

$$\chi^2 = \frac{(10-5)^2}{5} + \frac{(5-0)^2}{5} = 10$$

Critical value chart

http://stattrek.com/chi-square-test/goodness-of-fit.aspx?tutorial=ap

Connection between binomial, normal, chi square

- N trial, m success, p is success rate, q = 1 - p

	success	fail
observed	m	N-m
expected	Np	Nq

$$\chi^2 = \frac{(m-Np)^2}{(Np)} + \frac{(N-m-Nq)^2}{(Nq)}$$

$$\chi^2 = \frac{(m-Np)^2}{(Npq)}$$

$N = Np + N(1-p)$
$N = m + (N-m)$
$q = 1-p$

$$\chi = \frac{m-Np}{\sqrt{(Npq)}}$$

CLT

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

O_i = the number of observations of type *i*.
$E_i = N * p_i$, the expected frequency of type *i*,
n = the number of cells in the table.

m/N ~ N(p, pq/N)
Proof when multinomial case:
http://sites.stat.psu.edu/~dhunter/asymp/fall2006/lectures/ANGELchpt07.pdf

https://www.stat.washington.edu/peter/342/non_param.pdf

Chi square test (two sample)

- Test if X, Y follow same distribution
  - H_0: X and Y follow same distribution
    H_a: X and Y do not follow same distribution
  - Example
    - Two types of bidder, human and computer. Different categories of bidding product. Test computer bids similarly across categories as humans.

$$chi\ square = \sum_{all\ cells} \frac{(observed - expected)^2}{expected}$$

How to calculate the expected stats?

Chi square test (cont'd)

	Auto parts	Books music	cloth	computer	furniture	home goods	jewelry	mobile	Office equip	Sporting goods	**Row sum**
Human	9757	13733	476	9733	87807	389249	555634	492350	160671	939398	2658808
Robot	0	1509	0	11667	0	18708	37101	105138	7967	230326	412416
column sum	9757	15242	476	21400	87807	407957	592735	597488	168638	1169724	

X_1 → Human row; Y_1 → Robot row; Col sum_1 → column sum row; Row sum_1 → Human row total; Row sum_2 → Robot row total

- If same distribution
  - $X_1 : Y_1$ = row sum_1 : row sum_2
  - $X_1 + Y_1$ = col sum_1
  - Expected value X_1 = col sum_1 * row sum_1 / (row sum_1 + row sum_2)

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

df: (row num - 1) (column num - 1)

https://www.stat.berkeley.edu/~stark/SticiGui/Text/chiSquare.htm

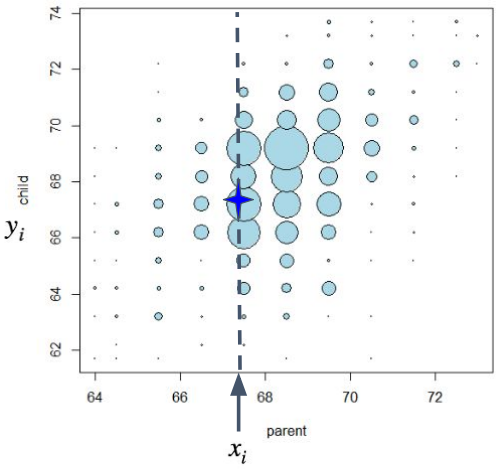
the multinomial distribution can be approximated by the theoretical chi square distribution under certain conditions. The condition ordinarily specified by statisticians is that the expected cases in each category be 5 or more

Summary

- Hypothesis testing (cont'd)
  - Test for variance, Chi square test
- Simple linear regression
  - Modeling E(Y|X)
  - Assumption
  - Coefficient estimation
    - Least square estimation
    - Maximum likelihood estimation
    - Hypothesis testing

Classical example

- Let y_i be the height of child i for i = 1, …, n; x_i be the height of child i's parent
- How to use x_i to predict y_i?
- What are we predicting?
  - E[Y | X = x]
- Simple linear regression
  - $E[Y | X = x] = \beta_0 + \beta X$



y_i changes due to a lot of reasons, x could only account for some of it. Therefore, for each specific x_i, the possible value of y_i actually has a range and the best guess we can perform is the center/expectation of y_i at certain x_i.

Simple linear regression model

- Assumptions
  - 1. Distribution of X is arbitrary
  - 2. Linear relationship: $Y = \beta_0 + \beta_1 X + \varepsilon$
  - 3. $E[\varepsilon \mid X = x] = 0$ $\quad Var[\varepsilon \mid X = x] = \sigma^2$ $\quad$ (no matter what x is)
  - 4. ε is independent: uncorrelated across observations
- **Gaussian-noise** assumption
  - 1, 2 are same
  - ~~Stronger assumption than 3, 4.~~ $\varepsilon \sim N(0, \sigma^2)$ and independent of X

Do we expect y to follow normal distribution?

Observations are independent, then we can have the independent error term assumption held.

http://www.stat.cmu.edu/~cshalizi/mreg/15/lectures/01/lecture-01.pdf
http://www.stat.cmu.edu/~cshalizi/mreg/15/lectures/04/lecture-04.pdf
error is normally distributed is the requirement, but response variable y doesn't have to be.
See explanation here:
http://stats.stackexchange.com/questions/12262/what-if-residuals-are-normally-distributed-but-y-is-not
http://stats.stackexchange.com/questions/11351/left-skewed-vs-symmetric-distribution-observed/11352#11352

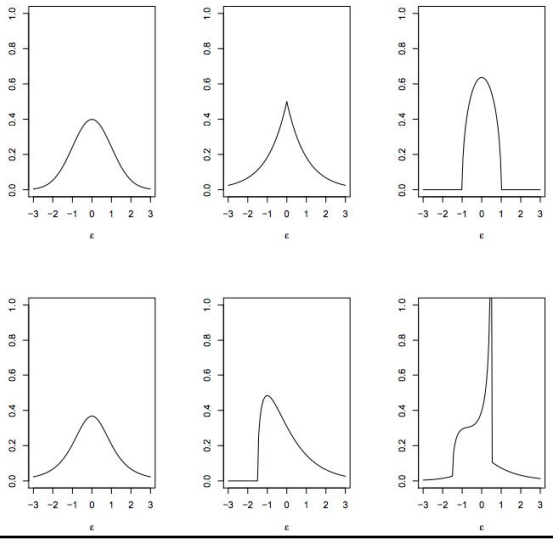
Demystify linear regression

- True relationship between X and Y might not be linear
  - But derived the optimal linear relationship approximation to the true one.
  - Why can we use linear to approximate? Taylor expansion
  - This approximation could be bad, but better than nothing.
- No assumption about distributions of X, Y or joint distribution of X and Y
- No assumption about causality that X causes Y
- No assumption that X is more precise, Y is more noisy
- It's not always normal distributed error term in the past

Since the distribution of X is arbitrary, distribution of Y is also arbitrary, since although there is noise in y, y's center is defined by x. Think of the example like most of x are close to 0, then most of y are close to beta0
https://stats.stackexchange.com/questions/12262/what-if-residuals-are-normally-distributed-but-y-is-not

https://stats.stackexchange.com/questions/152674/why-is-the-normality-of-residuals-barely-important-at-all-for-the-purpose-of-e

Some error term example



Why Gaussian noise?

- Central limit theorem
  - Noise might be sum of lots of little random noises from different sources, independent and with similar magnitude.
- Mathematical convenience
  - Closed form estimation

Coefficient estimation

- Some definition
  - Sale price *Y*: dependent variable, output, response
  - Lot area *X*: predictor, independent variable, covariate, input
- Coefficient estimation
  - Remember assumption $Y = f(X) + \varepsilon = \beta_0 + \beta_1 X + \varepsilon$
  - How do we find optimal (β_0, β_1), out of all possible (b_0, b_1)?
    - Least square error estimation (LSE)
    - Maximum likelihood estimation (MLE)
    - Compare LSE and MLE

Least square error estimation

- Mean square error (MSE)

$$
\begin{aligned}
MSE(b_0, b_1) &= \mathbb{E}\left[(Y - (b_0 + b_1 X))^2\right] \\
&= \mathbb{E}\left[Y^2\right] - 2b_0\mathbb{E}\left[Y\right] - 2b_1\mathbb{E}\left[XY\right] + \mathbb{E}\left[(b_0 + b_1 X)^2\right] \\
&= \mathbb{E}\left[Y^2\right] - 2b_0\mathbb{E}\left[Y\right] - 2b_1\mathrm{Cov}\left[X, Y\right] - 2b_1\mathbb{E}\left[X\right]\mathbb{E}\left[Y\right] + b_0^2 \\
&\quad +2b_0 b_1\mathbb{E}\left[X\right] + b_1^2\mathrm{Var}\left[X\right] + b_1^2(\mathbb{E}\left[X\right])^2
\end{aligned}
$$

E(X^2) = Var(X) + E(X)^2

$$
\frac{\partial\mathbb{E}\left[(Y - (b_0 + b_1 X))^2\right]}{\partial b_0} = -2\mathbb{E}\left[Y\right] + 2b_0 + 2b_1\mathbb{E}\left[X\right]
$$

$$
\begin{aligned}
\frac{\partial\mathbb{E}\left[(Y - (b_0 + b_1 X))^2\right]}{\partial b_1} &= -2\mathrm{Cov}\left[X, Y\right] - 2\mathbb{E}\left[X\right]\mathbb{E}\left[Y\right] + 2b_0\mathbb{E}\left[X\right] \\
&\quad +2b_1\mathrm{Var}\left[X\right] + 2b_1(\mathbb{E}\left[X\right])^2
\end{aligned}
$$

$$
\beta_1 = \mathrm{Cov}\left[X, Y\right]/\mathrm{Var}\left[X\right] \qquad \beta_0 = \mathbb{E}\left[Y\right] - \beta_1\mathbb{E}\left[X\right]
$$

$$
\Longrightarrow \qquad \hat{\beta}_1 = \frac{c_{XY}}{s_X^2} \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}
$$

population mean/variance and sample mean/variance, beta_1 and beta_0 are the true value if we have all the (x,y) data points, but since we only have a sample of (x,y), we have beta_1_hat and beta_0_hat, which are the estimators of true value beta_0 and beta_1.

Least square error estimation (cont'd)

- $\hat{\beta_0}, \hat{\beta_1}$ are estimators of β_0, β_1
  - $\hat{\beta}_1 = \frac{c_{XY}}{s_X^2} = \frac{cor(X,Y)\, s_Y}{s_X}$

    $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- Criteria for good estimators (week 1)
  - Unbiased

    $\mathbb{E}\left[\hat{\beta}_0\right] = \beta_0$

    $\mathbb{E}\left[\hat{\beta}_1\right] = \beta_1$

    How to prove?
  - Variance decrea[illegible] data (speed of convergence)

$$\mathrm{Var}\left[\hat{\beta}_1\right] = \frac{\sigma^2}{n s_X^2}$$

$$\mathrm{Var}\left[\hat{\beta}_0\right] = \frac{\sigma^2}{n}\left(1 + \frac{\bar{x}^2}{s_X^2}\right)$$

How to understand?

Are estimators unbiased?

- $$\begin{aligned}\hat{\beta}_1 &= \frac{c_{XY}}{s_X^2} \\ &= \frac{\frac{1}{n}\sum_{i=1}^n x_i y_i - \bar{x}\bar{y}}{s_X^2} \\ &= \frac{\frac{1}{n}\sum_{i=1}^n x_i(\beta_0 + \beta_1 x_i + \epsilon_1) - \bar{x}(\beta_0 + \beta_1\bar{x} + \bar{\epsilon})}{s_X^2} \\ &= \frac{\beta_0\bar{x} + \beta_1\overline{x^2} + \frac{1}{n}\sum_{i=1}^n x_i\epsilon_i - \bar{x}\beta_0 - \beta_1\bar{x}^2 - \bar{x}\bar{\epsilon}}{s_x^2} \\ &= \frac{\beta_1 s_X^2 + \frac{1}{n}\sum_{i=1}^n x_i\epsilon_i - \bar{x}\bar{\epsilon}}{s_X^2} \\ &= \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^n x_i\epsilon_i - \bar{x}\bar{\epsilon}}{s_X^2}\end{aligned}$$

$$\bar{x}\bar{\epsilon} = n^{-1}\sum_i \bar{x}\epsilon_i \qquad \hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})\epsilon_i}{s_X^2}$$

E(beta_1_hat) = beta_1 since E(error_i) = 0

Hypothesis testing for coefficient

- Test assumption regarding a (population) parameter β_1
  - Null hypothesis $\beta_1 = 0$ (What type of test)
- Calculate stats: compare observed and H_0
  - $\hat{\beta}_1 \sim N(\beta_1, \sigma^2/ns_X^2) \quad \hat{\beta}_0 \sim N(\beta_0, \frac{\sigma^2}{n}\left(1+\frac{\bar{x}^2}{s_X^2}\right))$

    $\widehat{\text{se}}\left[\hat{\beta}_1\right] = \frac{\hat{\sigma}}{s_x\sqrt{n}}$

    $\widehat{\text{se}}\left[\hat{\beta}_0\right] = \frac{\hat{\sigma}}{s_X\sqrt{n}}\sqrt{s_X^2+\bar{x}^2}$

    $\frac{\hat{\beta}_1-\beta_1}{\widehat{\text{se}}\left[\hat{\beta}_1\right]} \sim t_{n-2}$ $\frac{\hat{\beta}_0-\beta_0}{\widehat{\text{se}}\left[\hat{\beta}_0\right]} \sim t_{n-2}$
  - Why t distribution. Recall t score. How to decide df.

    $Z = \frac{\bar{X}-\mathrm{E}[X]}{\sigma(X)/\sqrt{n}} \Longrightarrow \frac{\bar{X}-\mu}{S/\sqrt{n}}$
- Type I, type II error, power?

http://stats.stackexchange.com/questions/117406/proof-that-the-coefficients-in-an-ols-model-follow-a-t-distribution-with-n-k-d

How to estimate the variance of the error term ? σ^2

Estimating for σ^2

- What is σ^2 ?
  - $\text{Var}[\epsilon|X=x] = \sigma^2 = \mathbb{E}[(Y-(\beta_0+\beta_1 X))^2]$
- In sample (empirical) estimator
  - $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{m}(x_i))^2$
- Criteria for good estimator: unbiased
  - $s^2 = \frac{n}{n-2}\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{m}(x_i))^2$
  - Intuition about (n - 2)

when multiple linear regression, how's the df changed?

Why do we care about value of sigma?

What is likelihood

- Probability
  - If we know $\beta_0 = 0$, $\beta_1 = 1$, possibility to observe $(x_1=70, y_1=72), \ldots, (x_n, y_n)$?
  - What about $(x_1=70, y_1=92)$?
- Reverse above logic
  - If we observe $(x_1=70, y_1=72), \ldots, (x_n, y_n)$
  - Likelihood (L) of $\beta_0 = 0$, $\beta_1 = 1$?
- Connection
  - $L(\beta_0 = b_0, \beta_1 = b_1 \mid \{(x_1, y_1), \ldots, (x_n, y_n)\}) = P(\{(x_1, y_1), \ldots, (x_n, y_n)\} \mid \beta_0 = b_0, \beta_1 = b_1)$
- How to use likelihood to estimate β_0, β_1
  - Maximizes $L(\beta_0 = b_0, \beta_1 = b_1 \mid \{(x_1, y_1), \ldots, (x_n, y_n)\})$

http://stats.stackexchange.com/questions/2641/what-is-the-difference-between-likelihood-and-probability
For a specific distribution, the value of likelihood and possibility is equal, but the meaning is not same.
About likelihood:
https://www.psychologicalscience.org/observer/bayes-for-beginners-probability-and-likelihood#.WR5_xlPyuLI

Maximum likelihood estimation (MLE)

- Under Gaussian error distribution

$$L(\beta_0, \beta_1 | \{(x_1, y_1), \ldots (x_n, y_n)\} = P(\{(x_1, y_1), \ldots (x_n, y_n)\} | \beta_0, \beta_1)$$

$$= \prod_{i=1}^{n} P((x = x_1, y = y_1 | \beta_0 = b_0, \beta_1 = b_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - (b_0 + b_1 x_i))^2}{2\sigma^2}}$$

$$log(L(\beta_0, \beta_1 | \{(x_1, y_1), \ldots (x_n, y_n)\}) = \sum_{i=1}^{n} log(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - (b_0 + b_1 x_i))^2}{2\sigma^2}})$$

$$= n\, log(\frac{1}{\sqrt{2\pi\sigma^2}}) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

Familiar?

- MLE to LSE are identical when $\varepsilon \sim N(0, \sigma^2)$

http://times.cs.uiuc.edu/course/410/note/mle.pdf

How to predict for new data?

- New data point x, how to predict y?
- Predicted (fitted) value at x: $\hat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$
- Is $\hat{m(x)}$ a single value?

$$E[\hat{m(x)}] = E[\hat{\beta_0} + \hat{\beta_1} x] = \beta_0 + \beta_1 x$$

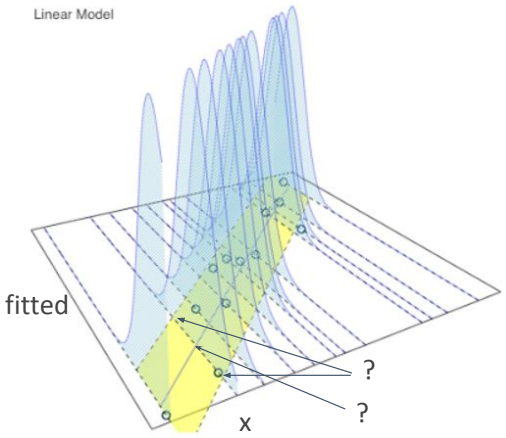
$$\text{Var}[\hat{m}(x)] = \frac{\sigma^2}{n}\left(1 + \frac{(x - \bar{x})^2}{s_X^2}\right)$$

  - Variance grows as σ^2 is larger; more noise in predictions
  - Larger n is, smaller variance is; more precise of predictions
  - First term in variance is var(y_bar), regardless of x.
  - Second term grows as (x - x_bar); further x is from center, less precise in prediction, and more spread out is x, more precise of predictions.

Distribution of predicted value

- When x = 70, 100% sure that y = 70? ($\hat{\beta}_0 = 0, \hat{\beta}_1 = 1$)
- Under Gaussian error assumption,

  $\hat{m}(x)$ follows normal distribution
  - What's Expectation, Variance, confidence interval?
- Visualization
  - Observations as circle
  - Mean prediction as solid line
  - [5th, 95th] quantile of predicted value.

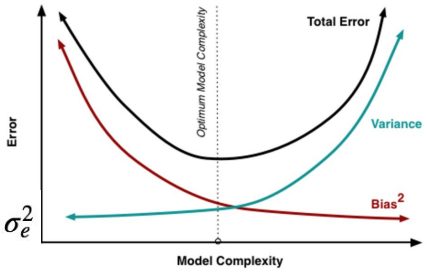


http://www.magesblog.com/2015/08/visualising-theoretical-distributions.html
question: there is variance in beta_1_hat and beta_0_hat, since y_hat = beta_0_hat + beta_1_hat * x, there is also variance in predicted value of y. If we talk about the 95% CI, that is predicted y, because true y also has the error term, which cannot be modeled.

Bias and variance tradeoff

- Prediction error, suppose we have $Y = f(X) + \varepsilon$

$$Err(x) = E\left[(Y - \hat{f}(x))^2\right]$$



$$Err(x) = \left(E[\hat{f}(x)] - f(x)\right)^2 + E\left[\left(\hat{f}(x) - E[\hat{f}(x)]\right)^2\right] + \sigma_e^2$$

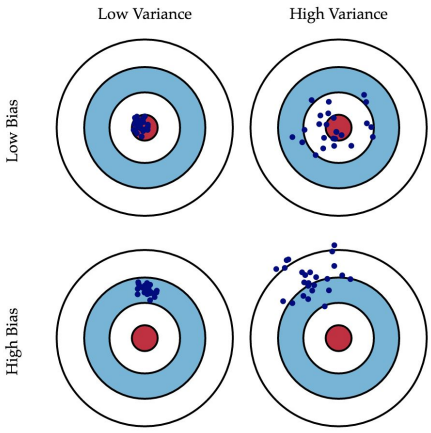
$$Err(x) = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff (proof)

http://scott.fortmann-roe.com/docs/BiasVariance.html

Bias and variance

- Accurate (small bias)
- Precise (small variance)



accuracy (mean) v.s precision (stderr): https://en.wikipedia.org/wiki/Accuracy_and_precision

Multiple linear regression

- More than one predictor, say p. *ith* data point

  $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \varepsilon_i$

- How to get estimate (LSE)?
  - MSE: 1/n * $\sum_{i=1}^{n} \left(Y_i - \sum_{k=0}^{p} X_{ki} \beta_j \right)^2$
  - (β_0 , β_1 , … β_p) minimizes MSE (derivative = 0).
- What about MLE? (HW)

LSE for multiple regression

- Matrix form
  - n*1 matrix ***Y***, n*(p+1) matrix ***X***, (p+1)*1 matrix β , n*1 matrix ε

    $Y = X\beta + \varepsilon$
  - MSE: $\frac{1}{n}(Y - X\beta)^T(Y - X\beta)$ $\nabla_\beta MSE(\beta) = \frac{2}{n}(-X^TY + X^TX\beta)$
  - Optimal estimator: $\hat{\beta} = (X^TX)^{-1}X^TY$
  - Fitted value: $\hat{Y} = X\hat{\beta} = X(X^TX)^{-1}X^TY$

Projection matrix: H

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

If students are not familiar with matrix, should check relative concepts like matrix calculation, like derivative, multiplication and stuff.

Property of estimator (beta)

- Recall simple linear regression
  - Bias?

$$\begin{aligned}\mathbb{E}\left[\widehat{\beta}|\mathbf{x}\right] &= \mathbb{E}\left[\beta + (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\epsilon|\mathbf{x}\right] \\ &= \beta + (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbb{E}\left[\epsilon|\mathbf{x}\right] \\ &= \beta\end{aligned}$$

  - Variance?

$$\begin{aligned}\mathrm{Var}\left[\widehat{\beta}|\mathbf{x}\right] &= \mathrm{Var}\left[\beta + (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\epsilon|\mathbf{x}\right] \\ &= \mathrm{Var}\left[(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\epsilon|\mathbf{x}\right] \\ &= (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathrm{Var}\left[\epsilon|\mathbf{x}\right]\mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1} \\ &= (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\sigma^2\mathbf{I}\mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1} \\ &= \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1} \\ &= \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}\end{aligned}$$

ANOVA

- ANOVA (Analysis of variance)
  - Decompose variance into sources
  - F test
    - Null hypothesis: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$
    - Calculation: (p does not include β_0)

Source of variance	df	Sum of square
Regression model	p	$sum((\hat{y} - \bar{y})^2)$
Residual	n-1-p	$sum((y - \hat{y})^2)$
Total	n-1	$sum((y - \bar{y})^2)$

$$F = \frac{Sum\ Square\ Regression\ /\ p}{Sum\ Square\ Residual\ /\ (n-p-1)}$$

F distribution

Proof
https://stats.stackexchange.com/questions/258461/proof-that-f-statistic-follows-f-distribution?noredirect=1&lq=1

F distribution is defined as chisq_var_1/df_1 / (chisq_var_2 / df_2)

Question: how to intuitively understand the sum of square follows chi square distribution.

Prove that sum square of total = sum square of model + sum square of residual, using sum of x_i * e_i = 0 and sum of e_i = 0. Here e_i is residual

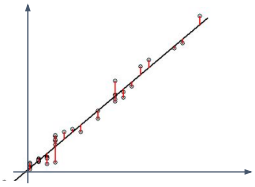
How to evaluate model performance?

- R^2, and relation with correlation $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
  - $$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \hat{\beta}_1^2 \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = cor(X, Y)^2$$
- Property
  - R^2, percentage of variance explained by model $$R^2 = \frac{SS_{model}}{SS_{model} + SS_{residual}}$$
  - $0 <= R^2 <= 1$
  - R^2 could be misleading, link
    - More features always increase R^2
    - Need to adjust R^2 $$R^2_{adj} = 1 - \left(\frac{n-1}{n-p-1}\right)(1 - R^2)$$

http://www.ats.ucla.edu/stat/mult_pkg/faq/general/noconstant.htm
https://www.r-bloggers.com/why-using-r-squared-is-a-bad-idea/
Why adding irrelevant features can still increase R_square, think of original feature space with dimension of p, then adding q more features, we have p+q dimension space. The best model in p dim space cannot be better than the best model in p+q dim space.

Residual vs. noise

- Residual $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$
- Noise $\epsilon_i = Y_i - (\beta_0 + \beta_1 x_i)$
  - e_i is a weighted sum of all ϵ_i



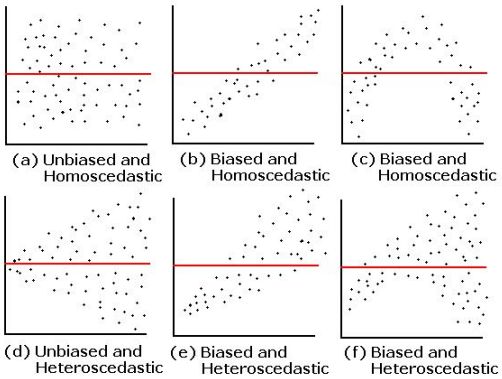
- Properties of residual
  - $\mathbb{E}[e_i|X = x] =$ 0
  - Constant variance, unchanging with x.
  - The residuals can't be completely uncorrelated with each other, but extremely weak, and grow negligible as n $\longrightarrow \infty$.
  - If the noise is Gaussian, the residuals is also Gaussian.

http://www.stat.cmu.edu/~cshalizi/mreg/15/lectures/07/lecture-07.pdf

Think one more time about the constant variance of residual, because when we need to standardize the residual, we know if x is farther away from center of x, the variance of residual is actually smaller, so that's why we need to standardize it.

Types of residuals

- Homoscedastic and Heteroscedastic



This is residual with X (or fitted value, which is just a linear transformation of X)

Residual diagnosis

- Recall the residual properties
- Understand the residual plots in R, link
  - Residual v.s. Fitted value: to check if unbiased and homoskedastic
  - QQ plot of residual: to check if errors are normally distributed
  - Square root of standardized residuals v.s. Fitted value: to check homoskedastic
    - What is standardization and why it is needed
  - Standardized residuals v.s. Leverage: to check outliers
    - Leverage: consider fitted line as lever, passing center point. Points further from center point have larger leverage.
      - $h_{ii} = [H]_{ii}$ measure distance of x to center of x

http://stackoverflow.com/questions/3505701/r-grouping-functions-sapply-vs-lapply-vs-apply-vs-tapply-vs-by-vs-aggrega
http://stats.stackexchange.com/questions/58141/interpreting-plot-lm (comprehensive answer)
http://strata.uga.edu/8370/rtips/regressionPlots.html (shorter answer to understand the plots)
https://en.wikipedia.org/wiki/Studentized_residual (why studentizing the residual)

standardized residuals, residuals divided by their standard deviations, since x is farther away from center, the residual is smaller. These points also have high leverage, or say high influential points, betas are adjusted for the sake of better fit of the high influential points. The leverage of ith data point is i * i element in projection matrix H, call it h_ii.

overall change in the coefficients when the ith point is deleted.

Leverage

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

$$\hat{y}_1 = h_{11}y_1 + h_{12}y_2 + \cdots + h_{1n}y_n$$
$$\hat{y}_2 = h_{21}y_1 + h_{22}y_2 + \cdots + h_{2n}y_n$$
$$\vdots$$
$$\hat{y}_n = h_{n1}y_1 + h_{n2}y_2 + \cdots + h_{nn}y_n$$

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \ldots + h_{ii}y_i + \ldots + h_{in}y_n \quad \text{for } i = 1, \ldots, n \Longrightarrow$$

h_{ii}: leverage of ith data point

- Properties of h_{ii}
  - h_{ii} is a measure of the distance between $\vec{x}_i$ and mean of the $\vec{x}$
  - h_{ii} is a number between 0 and 1.
  - Sum of the h_{ii} equals *p+1*, the number of parameters (regression coefficients including the intercept).

This slide is a result of how high leverage impact the prediction. So points with high leverage (larger value of h_ii), the predicted value of this point majorly depends on its own value y_i since there are not too many points around it. And also best values of betas are chosen in favor to fit these points.

Distinction Between Outliers and High Leverage Observations
https://onlinecourses.science.psu.edu/stat462/node/171
Outlier: unusual y value.
Influential point: unusual x value.
Outlier doesn't have to be deleted.
High influential point also doesn't have to be deleted.

Standardizing

- Standardized residual
  - What: Residuals rescaled to have a mean of 0 and a variance of 1
  - Why:
    - Since data points far from center has larger impact on estimating coefficient (leverage), residual variance of these data points are smaller

$$e = Y - \hat{Y} = Y - X\hat{\beta} = Y - X(X^TX)^{-1}X^TY = (I - H)Y$$

$$\mathrm{Var}(e) = \mathrm{Var}((I-H)Y) = (I-H)\,\mathrm{Var}(Y)(I-H)^\top = \sigma^2(I-H)^2 = \sigma^2(I-H)$$

$$\mathrm{Var}(e_i) = (1-h_{ii})\sigma^2$$

  - How to standardize: $t_i = \dfrac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$

Standardizing in simple linear regression

- $\mathrm{Var}(e_i) = (1 - h_{ii})\sigma^2$

$$X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

$$\mathrm{Var}(e_i) = \sigma^2 \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \right)$$

- Residual and leverage
  - Link

Should we standardize features?

- Standardizing won't change coefficient significance

$$\hat{\beta}_1(x_1) = \frac{\sum_{i=1}^{n}(x_{1,i} - \bar{x}_1)(y_i - \bar{y})}{\sum_{i=1}^{n}(x_{1,i} - \bar{x}_1)^2}$$

$$\hat{\beta}_1(ax_1) = \frac{\sum_{i=1}^{n}(ax_{1,i} - a\bar{x}_1)(y_i - \bar{y})}{\sum_{i=1}^{n}(ax_{1,i} - a\bar{x}_1)^2} = \frac{a\sum_{i=1}^{n}(x_{1,i} - \bar{x}_1)(y_i - \bar{y})}{a^2\sum_{i=1}^{n}(x_{1,i} - \bar{x}_1)^2} = \frac{\hat{\beta}_1(x_1)}{a}$$

- For comparing coefficients for different predictors within a model, standardizing helps
- For comparing coefficients for the same predictors across different data sets, don't standardize

http://stats.stackexchange.com/questions/29781/when-conducting-multiple-regression-when-should-you-center-your-predictor-varia

http://andrewgelman.com/2009/07/11/when_to_standar/

The benefit of scaling, for example, in regularization, we must need to scale.

Summary

- Hypothesis testing (cont'd)
  - Test for variance, Chi square test
- Simple linear regression, LSE, MLE
- Multiple linear regression, ANOVA
- Residual diagnosis