THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Online Review Course of
Undergraduate Probability and Statistics

# Review Lecture 17
# Linear Regression, part 2

Chris A. Mack
*Adjunct Associate Professor*

Course Website: www.lithoguru.com/scientist/statistics/review.html

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

## MLE Linear Regression

- Our model: $E[Y|X] = aX + b$

$$a = \frac{cov(X,Y)}{var(X)} \qquad b = E[Y] - aE[X]$$

- Our LSE estimators:

$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad b = \bar{y} - a\bar{x}$$

- For $Y_i$ normally distributed, the MLE is the same as the LSE

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

## MLE Linear Regression Assumptions

- Y is linearly related to X
- $X_i$ is known with certainty; only $Y_i$ has uncertainty, some of which is explained by $X_i$
- All $Y_i$ are iid normally distributed
- There are no outliers: our regression is not robust, and even one bad data point will mess up our results

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

## Linear Regression Steps

1. Create an x-y scatterplot of the data
   - Is the assumption of a linear relationship reasonable?
   - Are there any obvious outliers?
2. What do you know about the underlying pdfs of X and Y?
   - Do you have reason to assume normal distributions, or are you doing so for convenience?
3. Perform the regression
4. Plot the residuals
   - Check the assumption of iid normal

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

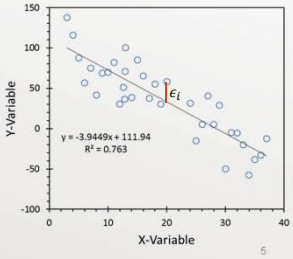
## Linear Regression Graphs

- Create an x-y scatterplot of the data

Model: $y_i = ax_i + b + \epsilon_i$

Predicted Value: $\hat{y}_i = ax_i + b$

Residual: $\epsilon_i = y_i - \hat{y}_i$



THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

## Linear Regression Graphs

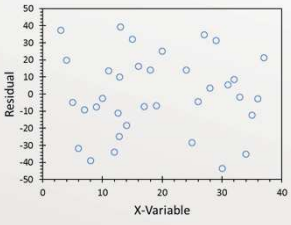
- Create an x-y scatterplot of the residuals

Plot $\epsilon_i$ vs. $\hat{y}_i$ *or* $x_i$

Plot a histogram of $\epsilon_i$

Look for violations of the iid normal assumption – do you see any trend in the residuals?

Note that $\bar{\epsilon} = 0$



THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

## Goodness of Fit

- Goodness of fit metric: $R^2 = r^2$

$$r = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} \qquad R^2 = \frac{cov^2(X,Y)}{var(X)var(Y)}$$

- Also, we can show that

$$R^2 = 1 - \frac{var(\epsilon)}{var(Y)}$$

$R^2$ is the fraction of the variance of Y that is explained by the linear fit

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

## Review #17: What have we learned?

- What are the assumptions for a standard least-squares linear regression?
- Why is it a good idea to plot our residuals?
- What is the goodness of fit and how do we interpret its value?
- How does one construct confidence intervals for the linear fit parameters and for predicted values from our fit equation?

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

## Uncertainty of Fit Parameters

- The regression fit is based on a sample of data

Population Model: $y_i = \alpha x_i + \beta + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma_\epsilon^2)$

- To create confidence intervals for $a, b, \hat{y}_i$, and $r$ we need to know their sampling distributions
  - They are unbiased and t-distributed (DF = n – 2)

$$var(a) = \frac{var(\epsilon)}{n\, var(X)} \qquad var(b) = \frac{var(\epsilon)}{n}\left(1 + \frac{\bar{x}^2}{var(X)}\right)$$

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

## Uncertainty of Predictions

- Uncertainty in $\hat{y}_i$ comes from the spread of the residuals and from uncertainty in the best fit parameters $a$ and $b$
- Again, the sampling distribution will be Student's t with DF = n – 2

$$var(\hat{y}_i) = \frac{var(\epsilon)}{n}\left(1 + \frac{(x_i - \bar{x})^2}{var(X)}\right)$$