A Primer on Mathematical Statistics and Univariate Distributions; The Normal Distribution; The GLM with the Normal Distribution

PSYC 943 (930): Fundamentals of Multivariate Modeling

Lecture 4: September 5, 2012

Today's Class

- The building blocks: The basics of mathematical statistics:
  - Random variables: definitions and types
  - Univariate distributions
    - General terminology
    - Univariate normal (aka, Gaussian)
    - Other popular (continuous) univariate distributions
  - Types of distributions: marginal, conditional, and joint
  - Expected values: means, variances, and the algebra of expectations
  - Linear combinations of random variables

- The finished product: How the GLM fits within statistics
  - The GLM with the normal distribution
  - The statistical assumptions of the GLM
  - How to assess these assumptions

RANDOM VARIABLES AND STATISTICAL DISTRIBUTIONS

Random Variables

Random: situations in which the certainty of the outcome is unknown and is at least in part due to chance

+

Variable: a value that may change given the scope of a given problem or set of operations

=

Random Variable: a variable whose outcome depends on chance

(possible values might represent the possible outcomes of a yet-to-be-performed experiment)

Today we will denote a random variable with a lower-cased: x

Types of Random Variables

- Random variables have different types:

1. Continuous
   - Examples of continuous random variables:
     - x represents the height of a person, drawn at random
     - Y_p (the outcome/DV in a GLM)

2. Discrete (also called categorical, generally)
   - Example of discrete:
     - x represents the gender of a person, drawn at random

3. Mixture of Continuous and Discrete:
   - Example of mixture:
     - x represents $\begin{cases} \text{response time (if between 0 and 45 seconds)} \\ 0 \end{cases}$

Key Features of Random Variables

- Random variables each are described by a **probability density/mass function (PDF)** $f(x)$ that indicates relative frequency of occurrence
  - A PDF is a mathematical function that gives a rough picture of the distribution from which a random variable is drawn

- The type of random variable dictates the name and nature of these functions:
  - Continuous random variables:
    - $f(x)$ is called a probability density function
    - Area under curve must equal 1 (found by calculus – integration)
    - Height of curve (the function value $f(x)$):
      - Can be any positive number
      - Reflects relative likelihood of an observation occurring
  - Discrete random variables:
    - $f(x)$ is called a probability mass function
    - Sum across all values must equal 1
    - The function value $f(x)$ is a probability (so must range from 0 to 1)

Other Key Terms

- The **sample space** is the set of all values that a random variable x can take:
  - The sample space for a random variable x from a normal distribution ($x \sim N(\mu_x, \sigma_x^2)$) is $(-\infty, \infty)$ (all real numbers)
  - The sample space for a random variable x representing the outcome of a coin flip is $\{H, T\}$
  - The sample space for a random variable x representing the outcome of a roll of a die is $\{1, 2, 3, 4, 5, 6\}$
- When using generalized models (discussed next week), the trick is to pick a distribution with a sample space that matches the range of values **obtainable** by data

Uses of Distributions in Data Analysis

- Statistical models make distributional assumptions on various parameters and/or parts of data

- These assumptions govern:
  - How models are estimated
  - How inferences are made
  - How missing data may be imputed

- If data do not follow an assumed distribution, inferences may be inaccurate
  - Sometimes a very big problem, other times not so much

- Therefore, it can be helpful to check distributional assumptions prior to (or while) running statistical analyses
  - We will do this at the end of class

CONTINUOUS UNIVARIATE DISTRIBUTIONS

Continuous Univariate Distributions

- To demonstrate how continuous distributions work and look, we will discuss three:
  - Uniform distribution
  - Normal distribution
  - Chi-square distribution
- Each are described a set of **parameters**, which we will later see are what give us our inferences when we analyze data
- What we then do is put constraints on those parameters based on hypothesized effects in data

Uniform Distribution

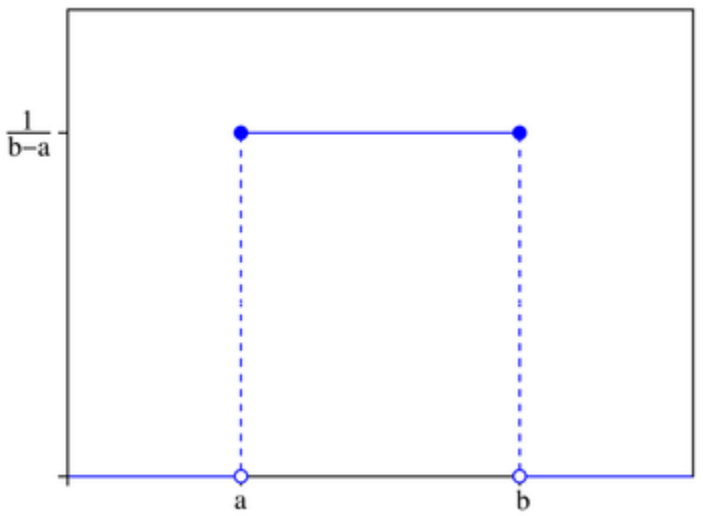
- The uniform distribution is shown to help set up how continuous distributions work

- For a continuous random variable x that ranges from (a, b), the uniform probability density function is:

$$f(x) = \frac{1}{b - a}$$

- The uniform distribution has two parameters:
  - a – the lower limit
  - b – the upper limit

- $x \sim U(a, b)$



More on the Uniform Distribution

- To demonstrate how PDFs work, we will try a few values:

x	a	b	$f(x)$
.5	0	1	$\frac{1}{1-0} = 1$
.75	0	1	$\frac{1}{1-0} = 1$
15	0	20	$\frac{1}{20-0} = .05$
15	10	20	$\frac{1}{20-10} = .1$

- The uniform PDF has the feature that all values of x are **equally likely** across the sample space of the distribution
  - Therefore, you do not see x in the PDF $f(x)$

- The mean of the uniform distribution is $\frac{1}{2}(a+b)$
- The variance of the uniform distribution is $\frac{1}{12}(b-a)^2$

Univariate Normal Distribution

- For a continuous random variable x (ranging from $-\infty$ to ∞) the univariate normal distribution function is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right)$$

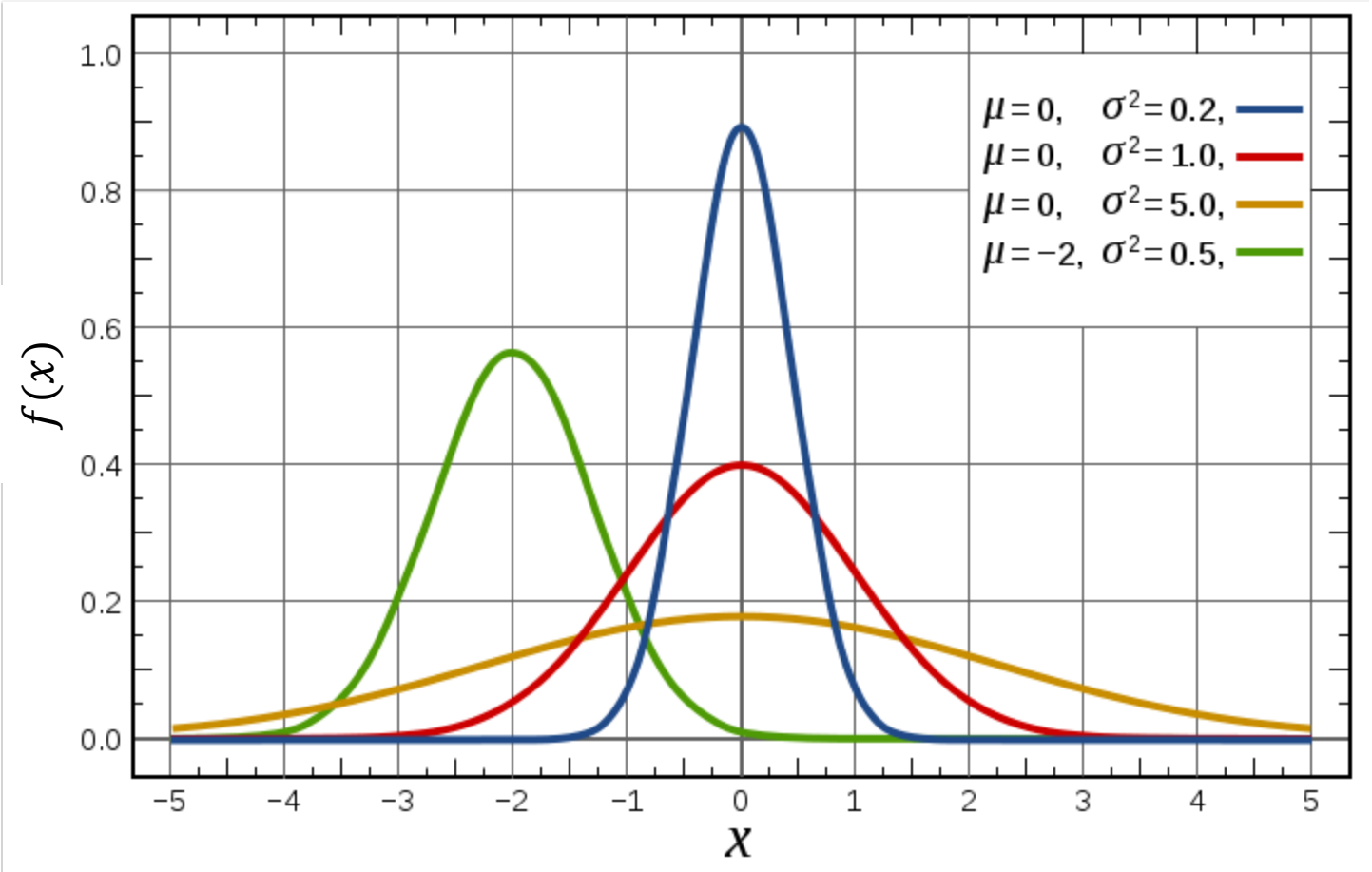
- The shape of the distribution is governed by two parameters:
  - The mean μ_x
  - The variance σ_x^2
  - These parameters are called **sufficient statistics** (they contain all the information about the distribution)

- The skewness (lean) and kurtosis (peakedness) are fixed

- Standard notation for normal distributions is $x \sim N(\mu_x, \sigma_x^2)$
  - Read as: "x follows a normal distribution with a mean μ_x and a variance σ_x^2"

- Linear combinations of random variables following normal distributions result in a random variable that is normally distributed
  - You'll see this later with respect to the GLM…

Univariate Normal Distribution



$f(x)$ gives the height of the curve (relative frequency) for any value of x, μ_x, and σ_x^2

More of the Univariate Normal Distribution

- To demonstrate how the normal distribution works, we will try a few values:

x	μ_x	σ_x^2	$f(x)$
.5	0	1	0.352
.75	0	1	0.301
.5	0	5	0.079
.75	-2	1	0.009
-2	-2	1	0.399

- The values from $f(x)$ were obtained by using Excel
  - The "=normdist()" function
  - Most statistics packages have a normal distribution function

- The mean of the normal distribution is μ_x
- The variance of the normal distribution is σ_x^2

Chi-Square Distribution

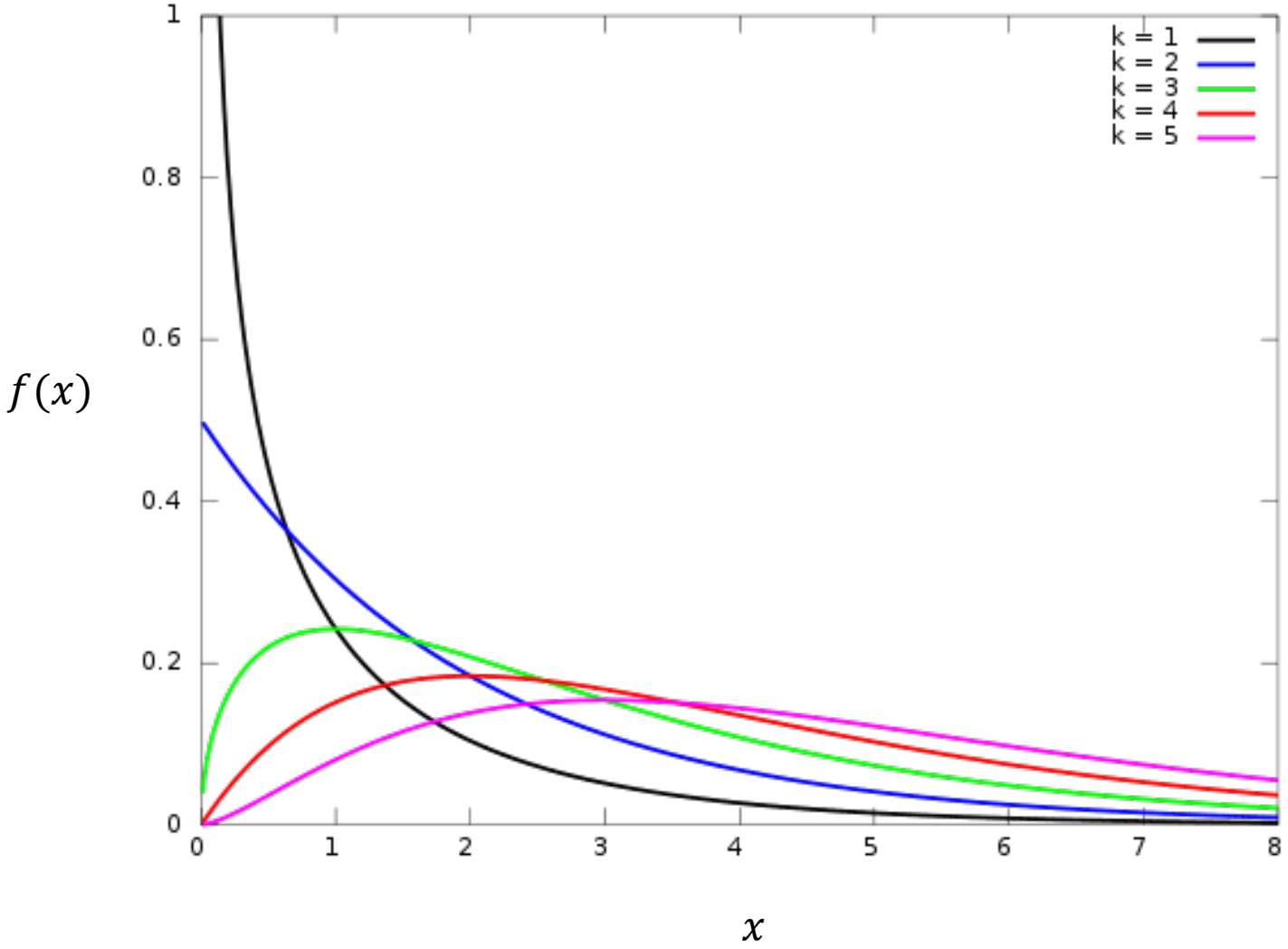
- Another frequently used univariate distribution is the Chi-Square distribution
  - Sampling distribution of the variance follows a chi-square distribution
  - Likelihood ratios follow a chi-square distribution

- For a continuous random variable x (ranging from 0 to ∞), the chi-square distribution is given by:

$$f(x) = \frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} x^{\frac{\nu}{2}-1} \exp\left(-\frac{x}{2}\right)$$

- $\Gamma(\cdot)$ is called the gamma function
- The chi-square distribution is governed by one parameter: ν (the degrees of freedom)
  - The mean is equal to ν; the variance is equal to 2ν

(Univariate) Chi-Square Distribution



MARGINAL, JOINT, AND CONDITIONAL DISTRIBUTIONS

Moving from One to Multiple Random Variables

- When more than one random variable is present, there are several different types of statistical distributions:

- We will first consider two discrete random variables:
  - x is the outcome of the flip of a penny (H_p, T_p)
    - $f(x = H_p) = .5$; $f(x = T_p) = .5$
  - z is the outcome of the flip of a dime (H_d, T_d)
    - $f(z = H_d) = .5$; $f(z = T_d) = .5$

- We will consider the following distributions:
  - Marginal distribution
    - The distribution of one variable only (either $f(x)$ ***or*** $f(z)$)
  - Joint distribution
    - $f(x, z)$: the distribution of both variables (both x ***and*** z)
  - Conditional distribution
    - The distribution of one variable, conditional on values of the other:
      - $f(x|z)$: the distribution of x given z
      - $f(z|x)$: the distribution of z given x

Marginal Distributions

- Marginal distributions are what we have worked with exclusively up to this point: they represent the distribution of one variable by itself
  - Continuous univariate distributions:
    - Uniform
    - Normal
    - Chi-square
  - Categorical distributions in our example:
    - The flip of a penny $f(x)$
    - The flip of a dime $f(z)$

Joint Distributions

- Joint distributions describe the distribution of more than one variable, simultaneously
  - Representations of multiple variables collected

- Commonly, the joint distribution function is denoted with all random variables separated by commas
  - In our example, $f(x, z)$ is the joint distribution of the outcome of flipping both a penny and a dime
    - As both are discrete, the joint distribution has four possible values:

$$f(x = H_p, z = H_d), f(x = H_p, z = T_d), f(x = T_p, z = H_d), f(x = T_p, z = T_d)$$

- Joint distributions are **multivariate distributions**
  - We will cover the continuous versions of these in a few weeks
    - The multivariate normal distribution
  - For our purposes, we will use joint distributions to introduce two topics
    - Joint distributions of independent variables
    - Joint likelihoods (next class) – used in maximum likelihood estimation

Joint Distributions of Independent Random Variables

- Random variables are said to be independent if the occurrence of one event makes it neither more nor less probable of another event
  - For joint distributions, this means: $f(x,z) = f(x)f(z)$

- In our example, flipping a penny and flipping a dime are independent – so we can complete the following table of their joint distribution:

Joint (Penny, Dime)

Dime

Penny	$z = H_d$	$z = T_d$	
$x = H_p$	$f(x = H_p, z = H_d)$	$f(x = H_p, z = T_d)$	$f(x = H_p)$
$x = T_p$	$f(x = T_p, z = H_d)$	$f(x = T_p, z = T_d)$	$f(x = T_d)$
	$f(z = H_d)$	$f(z = T_d)$	

Marginal (Penny)

Marginal (Dime)

Joint Distributions of Independent Random Variables

- Because the coin flips are independent, this becomes:

Dime

Joint (Penny, Dime)

Penny	$z = H_d$	$z = T_d$	
$x = H_p$	$f(x = H_p)f(z = H_d)$	$f(x = H_p)f(z = T_d)$	$f(x = H_p)$
$x = T_p$	$f(x = T_p)f(z = H_d)$	$f(x = T_p)f(z = T_d)$	$f(x = T_d)$
	$f(z = H_d)$	$f(z = T_d)$	

Marginal (Penny)

Marginal (Dime)

- Then, with numbers:

Dime

Joint (Penny, Dime)

Penny	$z = H_d$	$z = T_d$	
$x = H_p$	.25	.25	.5
$x = T_p$	.25	.25	.5
	.5	.5	

Marginal (Penny)

Marginal (Dime)

Marginalizing Across a Joint Distribution

- If you had a joint distribution, $f(x, z)$, but wanted the marginal distribution of either variable ($f(x)$ or $f(z)$) you would have to **marginalize** across one dimension of the joint distribution

- For categorical random variables, **marginalize = sum across**

$$f(x) = \sum_{z} f(x, z)$$

  - For example $f(x = H_p) = f(x = H_p, z = H_p) + f(x = H_p, z = T_p) = .5$

- For continuous random variables, **marginalize = integrate across**
  - No integration needed from you – just a conceptual understanding
  - Here, the integral = an eraser!

$$f(x) = \int_{z} f(x, z)dz$$

Conditional Distributions

- For two random variables x and z, a conditional distribution is written as: $f(z|x)$
  - The distribution of z given x

- The conditional distribution is also equal to the joint distribution divided by the marginal distribution of the conditioning random variable

$$f(z|x) = \frac{f(z,x)}{f(x)}$$

- Conditional distributions are found everywhere in statistics
  - As we will see, the general linear model uses the conditional distribution of the dependent variable (where the independent variables are the conditioning variables)

Conditional Distributions

- For discrete random variables, the conditional distribution is fairly easy to show:

Joint (Penny, Dime)

Dime

Penny	$z = H_d$	$z = T_d$	Marginal (Penny)
$x = H_p$	.25	.25	.5
$x = T_p$	.25	.25	.5
	.5	.5	

Marginal (Dime)

Conditional: $f(z|x = H_p)$:

$$f(z = H_d|x = H_p) = \frac{f(z = H_d, x = H_p)}{f(x = H_p)} = \frac{.25}{.5} = .5$$

$$f(z = T_d|x = H_p) = \frac{f(z = T_d, x = H_p)}{f(x = H_p)} = \frac{.25}{.5} = .5$$

We will show a continuous conditional distribution with the GLM in a few slides

EXPECTED VALUES AND THE ALGEBRA OF EXPECTATIONS

Expected Values

- Expected values are statistics taken the sample space of a random variable: they are essentially weighted averages

- The weights used in computing this average correspond to the probabilities (for a discrete random variable) or to the densities (for a continuous random variable).

- Notation: the expected value is represented by: $E(x)$
  - *The actual statistic that is being weighted by the PDF is put into the parentheses where x is now*

- Expected values allow us to understand what a statistical model implies about data, for instance:
  - How a GLM specifies the (conditional) mean and variance of a DV

Expected Value Calculation

- For discrete random variables, the expected value is found by:

$$E(x) = \sum_{x} xP(X = x)$$

- For example, the expected value of a roll of a die is:

$$E(x) = (1)\frac{1}{6} + (2)\frac{1}{6} + (3)\frac{1}{6} + (4)\frac{1}{6} + (5)\frac{1}{6} + (6)\frac{1}{6} = 3.5$$

- For continuous random variables, the expected value is found by:

$$E(x) = \int_{x} xf(x)dx$$

- We won't be calculating theoretical expected values with calculus…we use them only to see how models imply things about our data

Variance and Covariance…As Expected Values

- A distribution's theoretical variance can also be written as an expected value:

$$V(x) = E\big(x - E(x)\big)^2 = E(x - \mu_x)^2$$

  - This formula will help us understand predictions made GLMs and how that corresponds to statistical parameters we interpret

- For a roll of a die, the theoretical variance is:

$$V(x) = E(x - 3.5)^2 = \frac{1}{6}(1 - 3.5)^2 + \frac{1}{6}(2 - 3.5)^2 + \frac{1}{6}(3 - 3.5)^2 + \frac{1}{6}(4 - 3.5)^2 + \frac{1}{6}(5 - 3.5)^2 + \frac{1}{6}(6 - 3.5)^2 = 2.92$$

  - Likewise, the SD is then $\sqrt{2.92} = 1.71$

- Likewise, for a pair of random variables x and z, the covariance can be found from their joint distributions:

$$Cov(x, z) = E(xz) - E(x)E(z) = E(xz) - \mu_x\mu_z$$

LINEAR COMBINATIONS OF RANDOM VARIABLES

Linear Combinations of Random Variables

A **linear combination** is an expression constructed from a set of terms by multiplying each term by a constant and then adding the results

$$x = a_1 v_1 + a_2 v_2 + \cdots + a_n v_n$$

- The linear regression equation is a linear combination

- More generally, linear combinations of random variables have specific implications for the mean, variance, and possibly covariance of the new random variable

- As such, there are predicable ways in which the means, variances, and covariances change
  - These terms are called the algebra of expectations

- To guide us through this process, we will use the descriptive statistics from the height/weight/gender example from our 1st class

Descriptive Statistics for Height/Weight Data

Variable	Mean	SD	Variance
Height	67.9	7.44	55.358
Weight	183.4	56.383	3,179.095
Female	0.5	0.513	0.263

Diagonal: Variance

Above Diagonal: Covariance

Correlation /Covariance	Height	Weight	Female
Height	55.358	334.832	-2.263
Weight	.798	3,179.095	-27.632
Female	-.593	-.955	.263

Below Diagonal: Correlation

Algebra of Expectations

Here are some properties of expected values (true for any type of random variable): x and z are random variables, c and d constants

Sums of Constants:

$$E(x + c) = E(x) + c$$
$$V(x + c) = V(x)$$
$$Cov(x + c, z) = Cov(x, z)$$

Products of Constants:

$$E(cx) = cE(x)$$
$$V(cx) = c^2V(x)$$
$$Cov(cx, dz) = cdCov(x, z)$$

Sums of Random Variables:

$$E(cx + dz) = cE(x) + dE(z)$$
$$V(cx + dz) = c^2V(x) + d^2V(z) + 2cd\big(Cov(x, z)\big)$$

Examples for Algebra of Expectations

- Image you wanted to convert weight from pounds to kilograms (where 1 pound = 0.453 kg)

$$Weight_{kg} = .453 Weight_{lb}$$

- The mean (expected value) of weight in kg:

$$E\left(Weight_{kg}\right) = E(.453 Weight_{lb}) = .453 E(Weight_{lb}) = .453 \overline{Weight}_{lb}$$
$$= .453 * 183.4 = 83.08\text{kg}$$

- The variance of weight in kg:

$$V\left(Weight_{kg}\right) = V(.453 Weight_{lb}) = .453^2 V(Weight_{lb})$$
$$= .453^2 * 3{,}179.095 = 652.38 kg^2$$

- The covariance of weight in kg with height in inches:

$$Cov\left(Weight_{kg}, Height\right) = Cov(.453 Weight_{lb}, Height)$$
$$= .453 Cov(Weight_{lb}, Height) = .453 * 334.832$$
$$= 151.68\text{kg} * \text{inches}$$

Don’t Take My Word For It…

SAS syntax for transforming weight in a DATA step:

```
DATA htwt;
INPUT id Gender $ height weight ;
IF Gender = 'F' THEN female=1; IF Gender = 'M' THEN female=0;
heightMC = height-67.9;
weightKG = 0.453*weight;
```

SAS syntax for marginal descriptive statistics and covariances:

```
*NEW SAMPLE STATISTICS FOR WEIGHT;
PROC MEANS DATA=htwt MEAN VAR;
VAR weight weightKG;
RUN;

PROC CORR DATA=htwt COV;
VAR weight weightKG height;
RUN;
```

SAS output:

The MEANS Procedure

Variable	Mean	Variance
weight	183.4000000	3179.09
weightKG	83.0802000	652.3788519

Covariance Matrix, DF = 19

	weight	weightKG	height
weight	3179.094737	1440.129916	334.831579
weightKG	1440.129916	652.378852	151.678705
height	334.831579	151.678705	55.357895

Where We Use This…The Dreaded ESTIMATE Statement

- The ESTIMATE statement in SAS computes the expected value and standard error (square root of variance) for a new random variable
  - The new random variable is a linear combination of the original model parameters (the fixed effects)
  - The original model parameters are considered "random" here as their sampling distribution is used (assuming normal errors and a large N)

```
MODEL score = Dgroup2 Dgroup3 Dgroup4 experience4 enthusiasm
              Dgroup2*experience4  Dgroup3*experience4  Dgroup4*experience4 / SOLUTION;
ESTIMATE 'experience for mini' experience4 1 dgroup2*experience4 1;
```

$$Estimate = 1 * \beta_{experience4} + 1 * \beta_{G2*experience4}$$

- Where:
  - $\beta_{experience4}$ has mean $\widehat{\beta_{experience4}}$ and variance $se\left(\widehat{\beta_{experience4}}\right)^2$
  - $\beta_{G2*experience4}$has mean $\widehat{\beta_{G2*experience4}}$ and variance $se\left(\widehat{\beta_{G2*experience4}}\right)^2$
  - There exists a covariance between $\widehat{\beta_{experience4}}$ and $\widehat{\beta_{G2*experience4}}$
    - We'll call this $Cov\left(\widehat{\beta_{experience4}}, \widehat{\beta_{G2*experience4}}\right)$

More ESTIMATE Statement Fun

- So…if the estimates are:

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	75.49934727	0.38707620	195.05	<.0001
Dgroup2	-10.07267266	0.54896179	-18.35	<.0001
Dgroup3	4.17623925	0.54961852	7.60	<.0001
Dgroup4	-6.04195829	0.54912685	-11.00	<.0001
experience4	-0.38518388	0.29569936	-1.30	0.1943
enthusiasm	-5.00727782	0.18730609	-26.73	<.0001
Dgroup2*experience4	-0.63103823	0.39198136	-1.61	0.1091
Dgroup3*experience4	-0.10925920	0.41111045	-0.27	0.7907
Dgroup4*experience4	0.16959725	0.41917025	0.40	0.6862

  - And $\widehat{Cov}(\beta_{experience4}, \beta_{G2*experience4}) = -.08756$

…What is:

$$E(Estimate) = E(1 * \beta_{experience4} + 1 * \beta_{G2*experience4})$$
$$= 1 * E(\beta_{experience4}) + 1 * E(\beta_{G2*experience4}) = -.385 - .631 = -1.016$$
$$V(Estimate) = V(1 * \beta_{experience4} + 1 * \beta_{G2*experience4})$$
$$= 1^2 V(\beta_{experience4}) + 1^2 V(\beta_{G2*experience4}) + 2 * 1$$
$$* 1Cov(\beta_{experience4}, \beta_{G2*experience4}) =$$
$$.296^2 + .391^2 - 2 * .08756 = .0653$$

$$se(Estimate) = \sqrt{V(Estimate)} = .257$$

Parameter	Estimate	Standard Error	t Value	Pr > |t|
experience for mini	-1.0162221	0.25685951	-3.96	0.0001

THE GENERAL LINEAR MODEL WITH WHAT WE HAVE LEARNED TODAY

The General Linear Model, Revisited

- The general linear model for predicting Y from X and Z:

$$Y_p = \beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p + e_p$$

In terms of random variables, under the GLM:

- e_p is considered random: $e_p \sim N(0, \sigma_e^2)$

- Y_p is dependent on the linear combination of X_p, Z_p, and e_p

- The GLM provides a model for the **conditional distribution** of the dependent variable, where the conditioning variables are the independent variables: $f(Y_p | X_p, Z_p)$
  - There are no assumptions made about X_p and Z_p - they are constants
  - The regression slopes $\beta_0, \beta_1, \beta_2, \beta_3$ are constants that are said to be fixed at their values (hence, called fixed effects)

Combining the GLM with Expectations

- Using the algebra of expectations predicting Y from X and Z:

The expected value (mean) of $f(Y_p|X_p, Z_p)$:

$$\hat{Y}_p = E(Y_p) = E(\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p + e_p)$$

Constants

Random Variable with $E(e_p) = 0$

$$= \beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p + E(e_p)$$
$$= \beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p$$

The variance of $f(Y_p|X_p, Z_p)$:

$$V(Y_p) = V(\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p + e_p) = V(e_p) = \sigma_e^2$$

Distribution of $f(Y_p|X_p, Z_p)$

- We just found the mean (expected value) and variance implied by the GLM for the conditional distribution of Y_p given X_p and Z_p
- The next question: what is the distribution of $f(Y_p|X_p, Z_p)$?
- Linear combinations of random variables that are normally distributed result in variables that are normally distributed
- Because $e_p \sim N(0, \sigma_e^2)$ is the only random term in the GLM, the resulting conditional distribution of Y_p is normally distributed:

$$Y_p \sim N(\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p, \sigma_e^2)$$

Model for the means: from fixed effects; literally gives mean of $f(Y_p|X_p, Z_p)$

Model for the variances: from random effects; gives variance of $f(Y_p|X_p, Z_p)$

Examining What This Means in the Context of Data

- If you recall from our first lecture, the final model we decided to interpret: Model 5

$$W_p = \beta_0 + \beta_1\left(H_p - \bar{H}\right) + \beta_2 F_p + \beta_3\left(H_p - \bar{H}\right)F_p + e_p$$

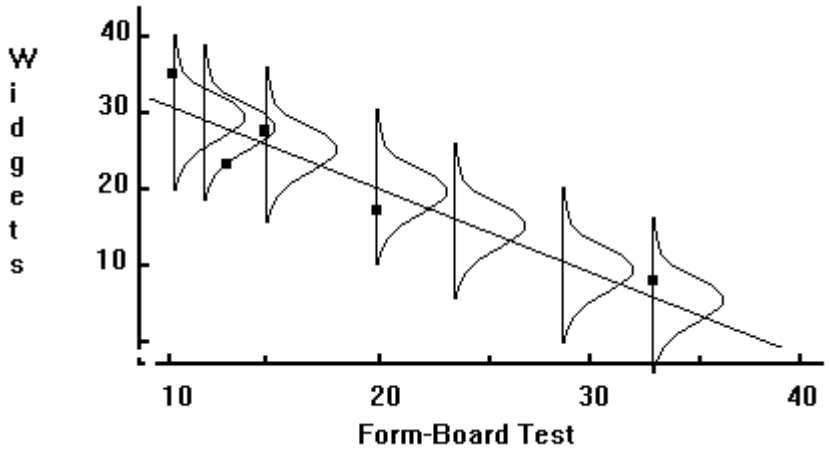
where $e_p \sim N(0, \sigma_e^2)$

- From SAS:

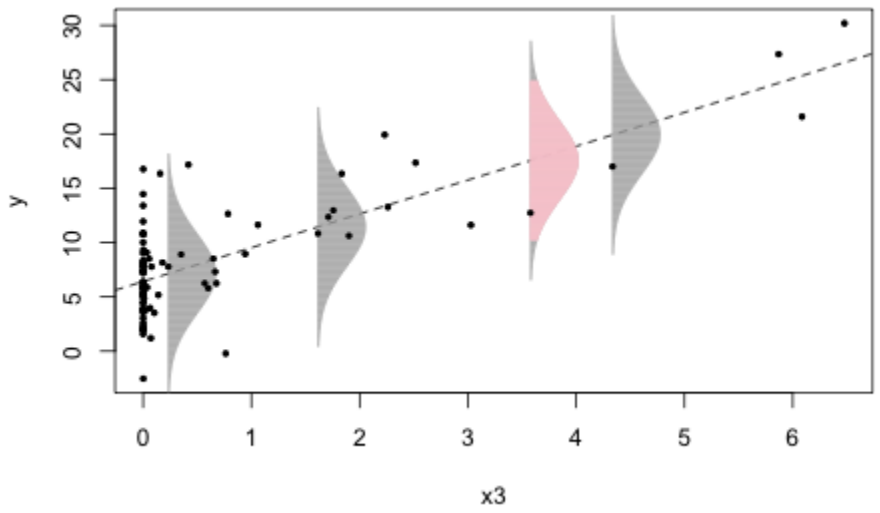
Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	222.1841719	0.83809108	265.11	<.0001
heightMC	3.1897275	0.11135027	28.65	<.0001
female	-82.2719216	1.21109969	-67.93	<.0001
heightMC*female	-1.0938553	0.16777741	-6.52	<.0001

Picturing the GLM with Distributions

The distributional assumptions of the GLM are the reason why we do not need to worry if our dependent variable is normally distributed



Our dependent variable should be **conditionally** normal

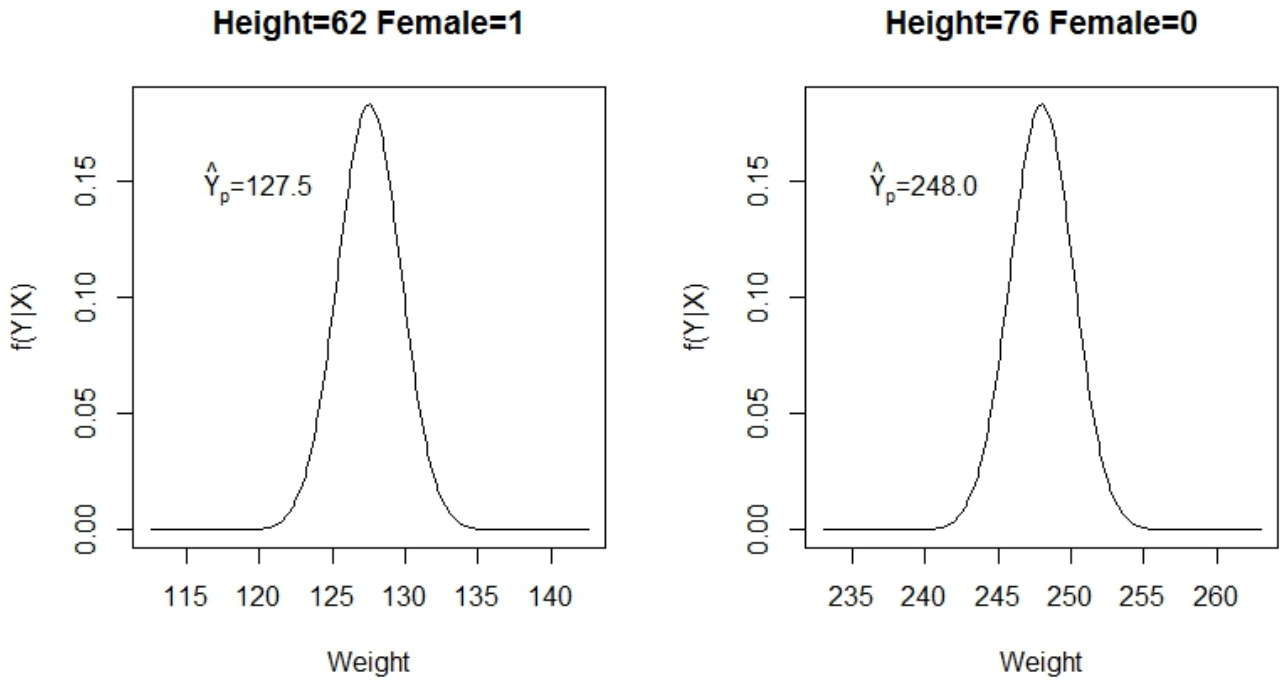


We can check this assumption by checking our assumption about the residuals, $e_p \sim N(0, \sigma_e^2)$

More on this soon…

More Pictures of the GLM

- Treating our estimated values of the slopes $(\beta_0, \beta_1, \beta_2, \beta_3)$ and the residual variance (σ_e^2) as the true values* we can now see what the theoretical* distribution of $f\left(Weight_p \middle| Height_p, Female_p\right)$ looks like for a given set of predictors



*Note: these distributions change when sample estimates are used (think standard error of the prediction)

Behind the Pictures…

- To emphasize the point that PDFs provide the height of the line, here is the normal PDF (with numbers) that produced those plots:

$$f(W_p|H_p,F_p)=\frac{1}{\sqrt{2\pi\sigma_e^2}}\exp\left(-\frac{(W_p-\widehat{W}_p)^2}{2\sigma_e^2}\right)$$

Model for the Means

$$=\frac{1}{\sqrt{2\pi\sigma_e^2}}\exp\left(-\frac{(W_p-(\beta_0+\beta_1(H_p-\bar{H})+\beta_2F_p+\beta_3(H_p-\bar{H})F_p))^2}{2\sigma_e^2}\right)$$

$$=\frac{1}{\sqrt{2\pi(4.73)}}\exp\left(-\frac{(W_p-(222.18+3.19(H_p-\bar{H})-82.27F_p-1.09(H_p-\bar{H})F_p))^2}{2(4.73)}\right)$$

Model for the Variance

The plots were created using the following value for the predictors:

$\bar{H}=67.9$

Left plot: $H_p=62; F_p=1$

Right plot: $H_p=76; F_p=0$

ASSESSING UNIVARIATE NORMALITY IN SAS

Assessing Univariate Normality in SAS

- The assumption of normally distributed residuals permeates GLM
  - Good news: of all the distributional assumptions, this seems to be the least damaging to violate. GLMs are robust to violations of normality.

- Methods exist to examine residuals from an analysis and thereby determine the adequacy of a model
  - Graphical methods: Quantile-Quantile plots (from PROC GLM)
  - Hypothesis tests (from PROC UNIVARIATE)

- Both approaches have problems
  - Graphical methods do not determine how much deviation is by chance
  - Hypothesis tests become overly sensitive to small deviations when sample size is large (have great power)

- To emphasize how distributions work, we will briefly discuss both

Assessing Distributional Assumptions Graphically

- A useful tool to evaluate the plausibility of a distributional assumption is that of the Quantile versus Quantile Plot (more commonly called a Q-Q plot)

- A Q-Q plot is formed by comparing the observed quantiles of a variable with that of a known statistical distribution
  - A quantile is the particular ordering of a given observation
  - In our data, a person with a height of 71 is the 39th tallest person (out of 50)
  - This would correspond to the person being at the $\frac{39-.5}{50} = .77$ or .77 percentile of the distribution (taller than 77% of the distribution)
  - The Q-Q plot then converts the percentile to a quantile using the sample mean and variance
    - A quantile is the value of an observation at the 77th percentile

- If the data deviate from a straight line, the data are not likely to follow from that theoretical distribution

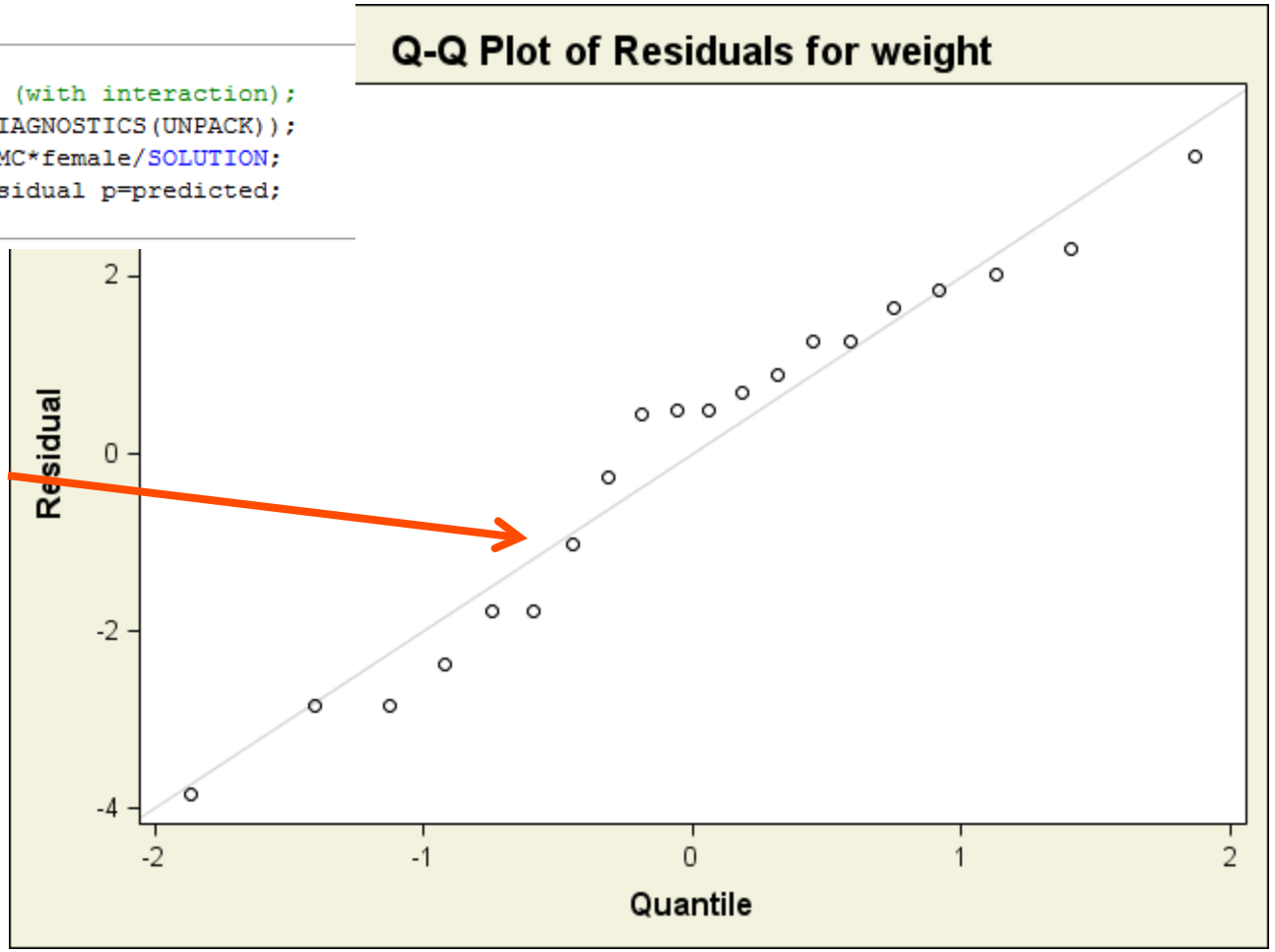
Q-Q Plots of GLM Residuals

- When you turn on the ODS Graphics, SAS PROC GLM will provide Q-Q plots for you with one command (PLOTS =):

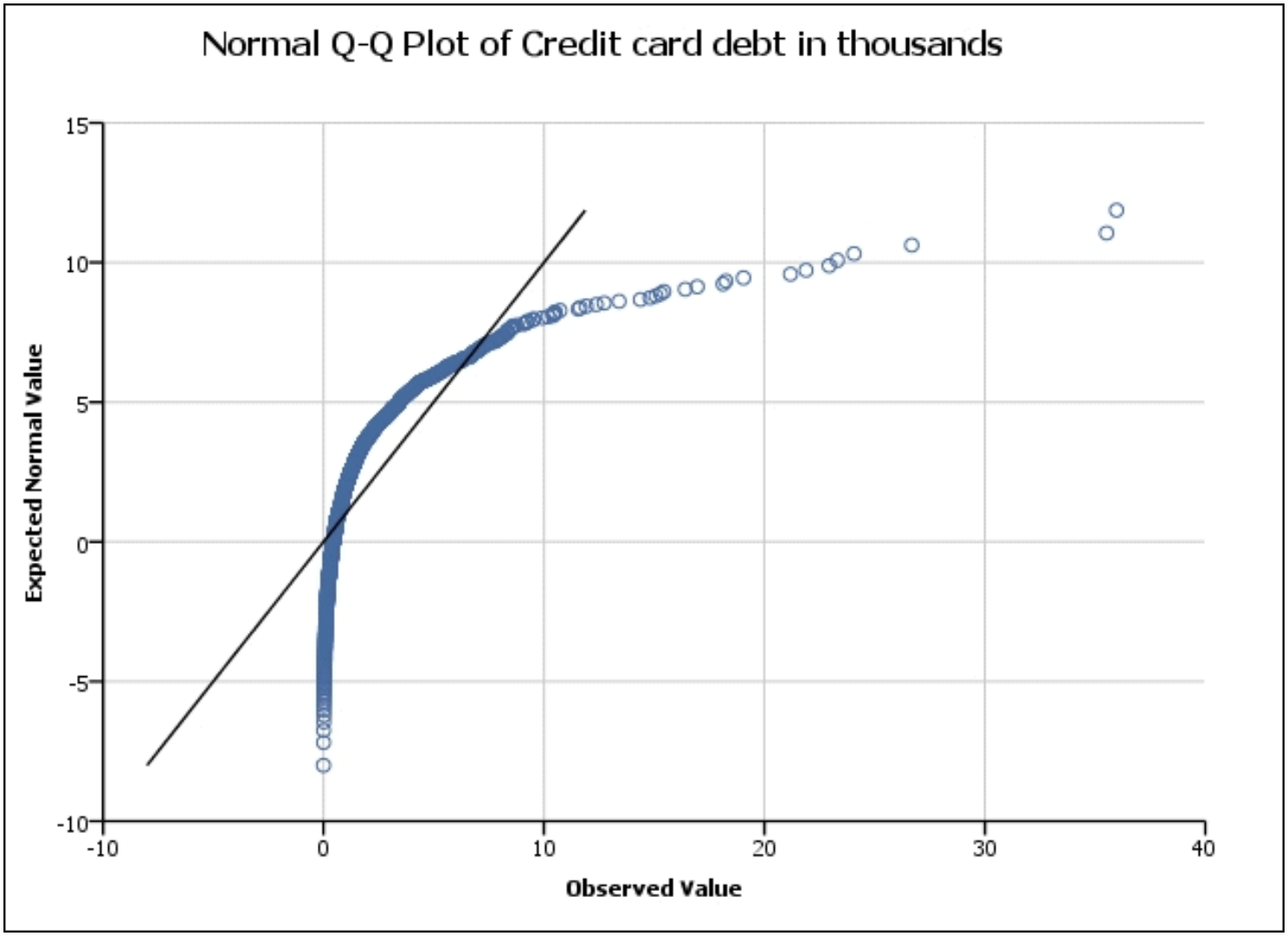
```
ODS HTML STYLE=harvest;
ODS GRAPHICS ON;

*Model #5: centered height and gender (with interaction);
PROC GLM DATA=work.htwt PLOTS= (ALL DIAGNOSTICS(UNPACK));
MODEL weight = heightMC female heightMC*female/SOLUTION;
OUTPUT OUT=work.diagnost residuals=residual p=predicted;
RUN; QUIT;
```

If residuals are normally distributed, they will fall on the line



Example Q-Q Plot of Non-Normal Data



Hypothesis Tests for Normality

- Additionally, using PROC UNIVARIATE, you can receive up to four hypothesis tests for testing H_0: Data come from normal distribution

```
*Model #5: centered height and gender (with interaction);
PROC GLM DATA=work.htwt PLOTS= (ALL DIAGNOSTICS(UNPACK));
MODEL weight = heightMC female heightMC*female/SOLUTION;
OUTPUT OUT=work.diagnost residuals=residual p=predicted;
RUN; QUIT;

PROC UNIVARIATE DATA=work.diagnost NORMAL;
VAR residual;
RUN;
```

Syntax for saving residuals in PROC GLM

Use new data set in PROC UNIVARIATE

PROC UNIVARIATE Output:

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	W	0.950554	Pr < W	0.3756
Kolmogorov-Smirnov	D	0.189844	Pr > D	0.0578
Cramer-von Mises	W-Sq	0.085849	Pr > W-Sq	0.1677
Anderson-Darling	A-Sq	0.461328	Pr > A-Sq	0.2377

If a given test is **significant**, then it is saying that your data **do not** come from a normal distribution

In practice, test will give diverging information quite frequently:
the best way to evaluate normality is to consider both plots and tests (approximate = good)

WRAPPING UP

Wrapping Up Today's Class

- Today was an introduction to mathematical statistics as a way to understand the implications statistical models make about data

- Although many of these topics do not seem directly relevant, they help provide insights that untrained analysts may not easily attain
  - They also help you to understand when and when not to use a model!

- We will use many of these same tools in our next class:

Estimation of GLMs:
Then (Least Squares) and Now (Maximum Likelihood)