Review of Regression

Joshua L Eubanks

University of Central Florida

Regression Equations

$$\hat{y} = \beta_0 + \beta_1 x_1 + \varepsilon$$

- $\hat{y}$ is the dependent variable, also referred to as *response* or *predicted* variable
- x is the independent variable, also referred to as *explanatory* or *predictor* variable
- β_0 is the intercept
- β_1 is the slope
- ε is the random disturbance, which is unknown

Analysis of Regression Equations

Prediction and forecasting (Thing 1: Plug and Chug): plug in new observations into regression equation and calculate the predicted value of the dependent variable.

Marginal Analysis (Thing 2): Multiply the change in one independent variable by its slope to find the predicted change in the dependent variable.

Trend Analysis

When time is the independent variable the equation becomes:

$$\hat{y} = \beta_0 + \beta_1 \mathit{Time} + \varepsilon$$

- Time trends: upward, downward, or no trend
- Trend rates only apply to time trends
- Trend rate is the coefficient of the time variable, or the slope regression equation (β_1)

Simple vs Multiple Regression

- Simple Regression:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \varepsilon$$

- Multiple Regression:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

  - $k =$ number of independent variables
  - So $k = 1$ for Simple Regression
  - $k > 1$ for Multiple Regression

Two Measures of Fit R^2 and SEE

- The regression is the best fitted equation, but how good is the best? To answer this we rely on two measures of fit:
  - $SEE = S =$ standard error (of the estimate), sometimes called the standard deviation of the regression
    - SEE is the Absolute Fit Measure that is in same units as the dependent variable
    - Thing 1 Prediction Margin of Error (MoE) is approximately $2 * SEE$ (for 95% cl)
  - $R^2 =$ R-Squared $=$ R-sqr
    - R^2 is the Relative measure of fit: $0\% \leq R^2 \leq 100\%$
    - Measure the percent of total variation in Y that is explained by the regression
    - Perfect Fit: $R^2 = 100\%$ all variation in Y is explained by the regression
    - No Fit: $R^2 = 0\%$ no variation in Y is explained by the regression

Adjusted R^2

- Comparing Different R^2's can be like comparing Apples to Oranges
  - with more variables (larger k) you have an edge
  - with less data (smaller n) it affects the comparison
  - To avoid comparing apples to Oranges, use the same data but with different variables (e.g. one model with say 5 variables and another with 6)
- Adjusted R^2 corrects for differences in k and n
  - Adjusted $R^2 < R^2$: the gap is roughly proportional to k
  - The larger the n, the smaller is that gap
- Adjusted R^2 compares different regressions
  - Adjusted R^2 is often reported instead of R^2

Sample Regression v. Population Regression

- Sample regression: fitted equation from a sample

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_k x_k$$

- Population regression equation: unknown relationship of dependent variable y with the k independent variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

Where $\beta_1, \beta_2, ..., \beta_k$ are population coefficients and ε is the random disturbance and the source of the unexplained variation in regression equation

- The expected value of the sample coefficient is the population coefficient: $\mathbb{E}(b_i) = \beta_i$

Summary of the Gauss Markov Theorem

Regression estimators are unbiased if the first 3 conditions are satisfied, and efficient if all five are satisfied:

- 1. parameters $\beta_0, \beta_1, \beta_2, ..., \beta_k$ have constant values
  - Lesson: inefficient estimates if not a homogeneous populations or time periods
- 2. expected value of random disturbance ε is zero
  - Lesson: intercept term helps out when $\mathbb{E}(\varepsilon)$ is far from zero
- 3. Independent variables are not correlated with ε
  - Lesson: Watch out for Specification Bias dangers (see below)
- 4. standard deviation of ε, σ_ε, remains constant
  - Lesson: inefficient cross section estimates if data varies widely
- 5. Uncorrelated ε's among different observations
  - Lesson: inefficient time series estimates if errors have a memory, which is referred to as autocorrelation (use Durbin Watson Statistic to test for autocorrelation)

Summary of the Gauss-Markov Theorem

Regression estimators are unbiased if the first 3 conditions are satisfied, and efficient if all five are satisfied. These are simple solutions to insure each condition is satisfied:

- 1. parameters have constant values
  - Solution: Study only homogeneous populations or time periods; use dummy variables to control for heterogeneity
- 2. expected value of random disturbance ε is zero
  - Solution: Always allow the software to calculate an intercept term (i.e. never force the slope coefficient to zero)
- 3. Independent variables are not correlated with ε
  - Solution: consider using ratios or percentages rather than raw data. Check for omitted variable bias, specification bias, multicollinearity

Summary Cont.

- 4. standard deviation of ε, σ_ε, remains constant
  - Solution: look for subgroups in data and analyze separately; use summary data (like the mean value) rather than the raw data
- 5. Uncorrelated ε's among different observations
  - Solution: transform the data into differences or percent change; use a more sophisticated time series model like the autoregressive model which will model the error term's rate of decay

Model Specification and Specification Bias

- Modeling: only represents the most important factors in a model
- Use independent variables that are the best proxies for each of those important factors
- Modeling errors are from omitting less important factors
- Random disturbance ε is sum of these errors (plus the sum of the measurement errors of variables we do include)
- Omitting less important variables won't affect significance test results for important variables we keep in the model
- Omitting important variables that should be in the model reduces your fit (lower R^2, higher SEE), and gives worse predictions
- But more importantly, you can no longer trust the t-test results if regression omits important independent variables that should be in model

Results of Specification Bias

- Biased coefficient estimates result from a misspecification of the regression model
- Specification bias may cause any of these problems:
  - Coefficient biased high or low by an unknown amount
  - Coefficient is the wrong sign (i.e., bias so large that it flips the sign of the coefficient)
  - Coefficient test gives wrong results: tests significant when it shouldn't, or doesn't test significant when it should

The Opposite of an Under-Specified model is an Over-Specified model

- Multicollinearity: some independent variables contain some of the same information
- Having some of the same information results in correlation between two or more independent variables
- Common amongst related proxy variables
- Common in time series analysis as variables tend to move together

Extreme Case of Multicollinearity and Violation of Gauss-Markov Assumption 3

- Can express one independent variable as linear combination of rest:amount of information $<$ number of indep variables
- e.g. Height in inches and feet (HtInches $=$ 12 HtFt) - Wt $=$ -380 $+$ 96 HtFt or Wt $=$ -380 $+$ 8 HtInch - HtFt $=$ 5, Wt $=$ 100 and HtInch $=$ 60, Wt $=$ 100
- Perfect collinearity: include both by allocating coefficients
  - Wt $=$ -380 $+$ 8 HtInch $+$ 0 HtFt
  - Wt $=$ -380 $+$ 0 HtInch $+$ 96 HtFt
  - Wt $=$ -380 $+$ 4 HtInch $+$ 48 HtFt
  - Wt $=$ -380 $+$ 16 HtInch – 96 HtFt
  - Wt $=$ -380 $+$ 88 HtInch – 960 HtFt
- Infinite number of equivalent versions, choice arbitrary

Perfect Multicollinearity

- Inflated coefficients (average of $\pm\infty$)
- Explanation: too many variables for amount of information in the data
- Result: Arbitrary regression coefficients, least squares explodes (computational singularity)
- Regression software saves us, fortunately
  - deletes last independent variable in collinear list
  - Equivalent of "Hey, dummy, you're repeating yourself."
- Treatment: avoid or remove redundant information from your model, e.g. NetWorth = Assets – Debt

Garden Variety Multicollinearity

- When two (or more) independent variable(s) is nearly a linear combination of the rest
- Simplest situation is if two independent variables are highly correlated
- In effect, the two independent variables provide approximately the same information
- More complex, general case is a subset of 3+ independent variables almost linearly related
- Warning: pairwise correlations may not be large when more than 2 variables involved

Problems with Severe Multicollinearity

- Least squares can't disentangle which regressors explain what
- Potential Damage of Multicollinearity:
  - inflated standard errors of regression coefficients, reduced t-ratios, high p-values
  - regression coefficients: arbitrary, volatile, unreliable
  - important regressors may not test significant

Can't Blame The Following Problems on Multicollinearity

- Entire model doesn't test significant (F test)
- Model has a poor fit (low R^2)
- Inaccurate and imprecise predictions or forecasts from the fitted regression equation
- Failure to test significant of regressors that are not part of the subset of collinear variables

Alternative Explanations that share similar symptoms

- Obvious alternative if not statistical significance: those regressors really do not matter
- Unjustified to blame multicollinearity for nonsignificance if correlations and VIFs not high
- VIF stands for variance inflation factor, which quantifies the severity of multicollinearity in an ordinary least squares regression analysis. VIF=1 no multicollinearity; 1<VIF<5 multicollinearity not a concern; VIF>5 multicollinearity "watch", VIF>10 multicollinearity "warning"
- VIF is sensitive to the number of independent variables. With small number of independent variables a VIF $= 5$ indicates a potential problem, whereas a VIF $= 5$ with a large number of independent variables does not.

Multicollinearity is not a problem unless it is a problem

- Not damaging if independent variables test significant despite high correlation
  - Just because correlations are high does not necessarily mean there is damage to test results
- Not damaging if primary goal is Thing 1 prediction or forecasting, multicollinearity not an issue
- Not damaging if collinearity is confined to "control variables" you're not planning to test individually anyway

How to Treat Multicollinearity

- Avoid poorly defined variables and poorly designed model specification:
  - Avoid using raw data: use ratios or percents instead of totals and subtotals
    - Ex.: include population and %Elderly, not population and #Elderly
  - Make default a common category to minimize correlations
    - Ex.: use majority population as your default, so in Orlando white is the default and %Latinos, %Black, and %Asian are included in the model

What not to do about Multicollinearity

Do not remove legitimately specified regressors

- Omitting important regressors risks specification bias and reduces overall fit
  - "Out of frying pan, into the fire"
  - Resulting bias makes test results questionable or worthless

Fear of Multicollinearity Led to Misspecified Model that Omits Number of Times Ranked (Top25)

Correlation Table of independent variables

	Win%10Yr	Enrollmt
Enrollmt	0.280	
Top25	0.663	0.287

AttendAve = - 60 + 1.5 Win%10Yr + 0.64 Enrollmt

Predictor	Coef	SE Coef	T	P
Constant	-59.75	12.26	-4.88	0.000
Win%10Yr	1.5245	0.2247	6.78	0.000←signif
Enrollmt	0.6438	0.1625	3.96	0.000

S = 16.1260 R-Sq = 48.4% R-Sq(adj) = 47.2%

Significance of Win Pct. Vanishes and Fit Way Up If Top25 Added in Correct Specification

The regression equation is

AttendAv = 0.80 + 0.29 Win%10Yr + 0.49 Enrollmt
+ 4.7 Top25CNN

Predictor	Coef	SE Coef	T	P	
Constant	0.80	10.71	0.07	0.941	
Win%10Yr	0.2874	0.2046	1.40	0.164	← not
Enrollmt	0.4876	0.1157	4.22	0.000	
Top25CNN	4.7376	0.4960	9.55	0.000	

S = 11.3634 R-Sq = 74.6% R-Sq(adj) = 73.8%

Bad Gas Price Model Omits Dow Index (DJIA): fear high Correlation with Other Variables

Correlations among independent variables

	P CRUDE	OPER%	PPI
OPER%	-0.501		
PPI	0.076	0.595	
DJIA	-0.396	0.765	0.700

P GAS = - 2.0 + 2.2 P CRUDE + 0.26 OPER% + 0.34 PPI

Predictor	Coef	SE Coef	T	P
Constant	-1.97	14.74	-0.13	0.894
P CRUDE	2.2414	0.1700	13.19	0.000
OPER%	0.2596	0.2305	1.13	0.264<-not signif
PPI	0.3395	0.1611	2.11	0.039

S = 4.65032 R-Sq = 82.5% R-Sq(adj) = 81.8%

Dow Improves Fit, Avoids Specification Bias: Oper% and PPI proxy for Dow's Inverse Effect on Gas Prices

The regression equation is

P GAS = - 46 + 2.0 P CRUDE + 0.53 OPER%
+ 0.74 PPI - 0.0077 DJIA

Predictor	Coef	SE Coef	T	P
Constant	-46.14	17.58	-2.62	0.011
P CRUDE	1.9994	0.1668	11.99	0.000
OPER%	0.5268	0.2207	2.39	0.020
PPI	0.7413	0.1794	4.13	0.000
DJIA	-0.007715	0.001985	-3.89	0.000

S = 4.23 R-Sq = 85.7% R-Sq(adj) = 84.9%

Wait on Mortgage Loan Approval: Bad Fit, Low Correlations, cannot blame Multicollinearity

Wait on Mortgage Loan Approval: Bad Fit, Low Correlations, cannot blame Multicollinearity

Correlations among independent variables NONE HIGH!

	INT	INC	LOAN
INC	-0.116		
LOAN	-0.235	0.347	
LATE30D	-0.035	0.299	0.054

WAIT = 7.8 - 0.36 INT + 0.018 INC - 0.026 LOAN + 0.19 LATE30D

Predictor	Coef	SE Coef	T	P
Constant	7.838	3.140	2.50	0.016
INT	-0.3571	0.2986	-1.20	0.237
INC	0.01816	0.02399	0.76	0.453
LOAN	-0.02630	0.01856	-1.42	0.162
LATE30D	0.1898	0.1053	1.80	0.077

S = 3.54912 R-Sq = 12.3% R-Sq(adj) = 5.7%

Summary of Practical Model Design: Basic Principles and Case Applications

- Include intercept so fit not forced thru origin: (Assump. 2)
- Don't mix apples with oranges in same model (Assump. 1)
- Avoid specification bias, wrong/suspect results (Assump. 3)
  - Include all "important" independent variables even if not interested in their effect or you fear multicollinearity damage
- Multicollinearity, at worst, damages only significance tests
  - For low correlations among indep. vars., no damage to tests
  - Causes no damage to forecasts, fit, or to uncorrelated variables
  - Only damaging if damaging: if variables don't test significant

Bad Modeling Practices

- Discard Variables Simply Because They Aren't Significant
  - Not significant: means p not $< \alpha$, but variable may still matter
    - Ex., p $= 0.30$ still is a 70% chance the variable is related
  - Variables are in model because they belong to control for their effects
  - May not test significant due to multicollinearity
- Torturing Data: Try All Possible Models and Number of Observations Until You Get the "Best" Model (i.e. p-hacking)
  - "Data Dredging" to find maximum fit often called unethical!
  - Data mining finds sample patterns of individuals that may not be present in population
  - Variables are in the model because they belong there to control for effects
    - Clients won't believe results if their favorite variable isn't in the model

Functional Forms

- Misspecification may occur even if you don't omit important variables
- Another type of misspecification is assuming a linear function when it's actually nonlinear

Tricking Linear regression into running nonlinear functions

Principle: CLS model (Classical Least Square) is linear in coefficients, not in variables

- Create (transform) data columns as new variables from other columns -e.g. quadratic, reciprocal, square root, log or ln (natural logs), product/interactive, ratio
- multiplicative production & demand functions transformed by natural logs
- If can't transform nonlinear to least squares, must use maximum likelihood estimation instead

Functional Forms and Marginal Change

Table 7.1 Summary of Alternative Functional Forms

Functional Form	Equation (one X only)	The Change in Y when X Changes
Linear	$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$	If X increases by one unit, Y will change by β_1 units.
Double-log	$\ln Y_i = \beta_0 + \beta_1 \ln X_i + \epsilon_i$	If X increases by one percent, Y will change by β_1 percent. (Thus β_1 is the elasticity of Y with respect to X.)
Semilog (lnX)	$Y_i = \beta_0 + \beta_1 \ln X_i + \epsilon_i$	If X increases by one percent, Y will change by $\beta_1/100$ units.
Semilog (lnY)	$\ln Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$	If X increases by one unit, Y will change by roughly $100\beta_1$ percent.
Polynomial	$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$	If X increases by one unit, Y will change by $(\beta_1 + 2\beta_2 X)$ units.

Consider Non-Linear Function if Justified by Theory, Experience, Logic, Previous Studies

1: Using Log of a variable - E.g. Population, enrollment, sales vary too much - E.g. Income directly related to education, but linear not logical (impact less at higher incomes)

$$Educ = \beta_0 + \beta_1 FamInc + \varepsilon$$

- So regress with a log-linear functional form

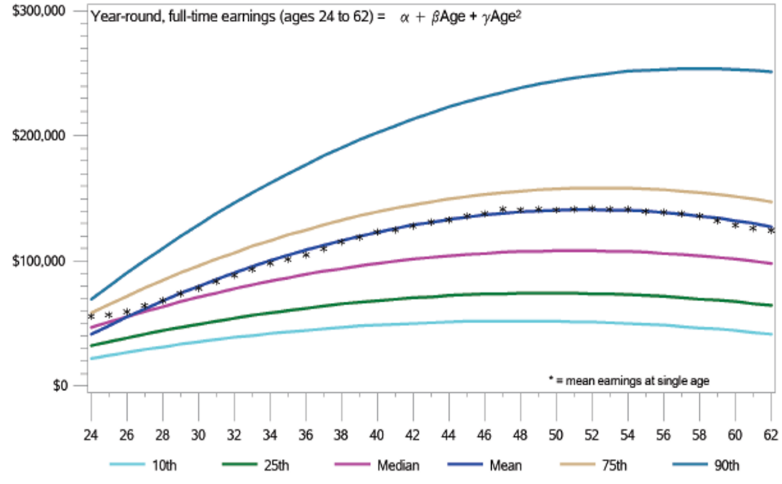
$$Educ = \beta_0 + \beta_1 Log10FamInc + \varepsilon$$

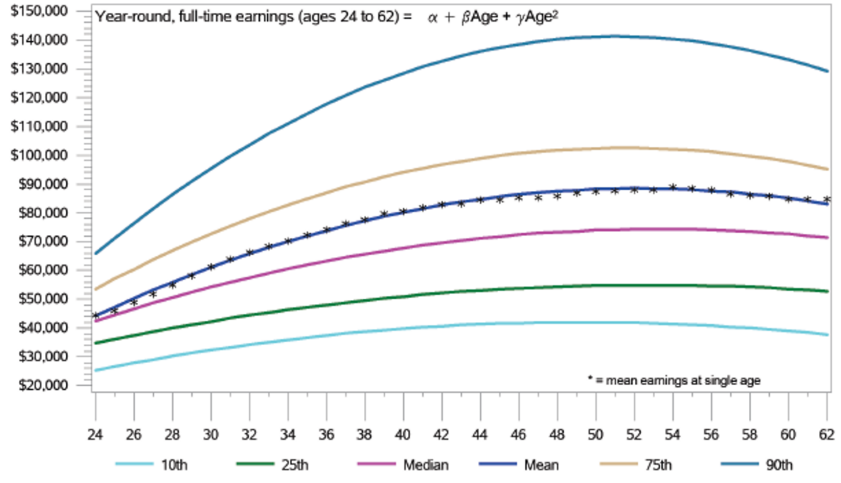
- Where define $Log10FamInc = \log_{10}(FamInc)$

Specify Quadratic if Expecting Maximum or Minimum

- Independent variable with too much of good thing $\rightarrow$ maximum
- E.g. Extra *Adv* helps less as the advertising budget increases (i.e. the law of diminishing returns):
  - NetInc $= \beta_0 + \beta_1 Adv + \beta_2 AdvSq + \varepsilon$
  - Where *AdvSq* is a transformed variable: $AdvSq = Adv^2$
- If both Adv and AdvSq test significant, then NetInc is a maximum at $Adv = -\frac{b_1}{2b_2}$
- Marginal change (thing 2) for $\Delta Adv = 1$ is: $\Delta NetInc = (b_1 + 2b_2 Adv)$
  - So, if $b_1 = 40, b_2 = -0.25$,
    - then max *NetInc* at $Adv = (-40/-0.5) = 80$
    - so at max $\Delta NetInc = (40 + 2 * -.25 * 80) = 0$
- What if your Adv budget is 60 and you want to ask for more?
  - so at each additional dollar to Adv buget causes NetInc to change by
    - $\Delta NetInc = (40 + 2 * -.25 * 60) = 10$

Case Study Nonlinear Specified as Quadratic: Age Earnings Profile, Master degree





Conforming age-earnings regression results

Level	α	β	γ	α error	β error	γ error	α t-stat	β t-stat	γ t-stat
10th	-67,908	5,015.47	-52.449	2,605	123.66	1.423	-26.1	40.6	-36.9
25th	-83,566	6,360.33	-64.092	2,511	123.94	1.460	-33.3	51.3	-43.9
Median	-111,514	8,623.02	-84.559	2,711	134.10	1.573	-41.1	64.3	-53.8
Mean	-205,739	13,401.87	-129.470	4,376	207.60	2.371	-47.0	64.6	-54.6
75th	-181,429	12,900.43	-123.312	3,001	180.24	2.220	-49.0	68.0	-53.3
90th	-286,486	18,711.63	-161.965	11,490	631.95	8.150	-24.9	29.6	-19.9

Cross-tabulation of earnings by age group

Ages	U.S. Pop	ACS Resp	10th	25th	Median	Mean	75th	90th
18 & over	3,561,045	182,583	42,479	62,694	93,239	118,808	134,678	205,851
18 to 62	3,412,525	173,973	42,853	62,656	93,239	118,451	134,678	203,813
18 to 24	18,788	866	20,379	30,878	50,559	53,548	69,028	87,487
18 to 19		n/d	n/d	n/d	n/d	n/d	n/d	n/d
20 to 24	18,788	866	20,379	30,878	50,559	53,548	69,028	87,487
25 to 34	797,338	37,586	34,491	49,913	70,783	81,088	97,779	130,649
25 to 29	290,528	13,316	29,204	42,470	60,671	66,589	82,341	104,520
30 to 34	300,810	24,270	40,141	53,871	77,089	89,399	104,320	143,007
35 to 44	1,090,648	53,690	47,789	66,303	96,063	119,445	134,678	201,774
35 to 39	548,751	26,672	46,317	62,712	91,007	110,358	125,423	180,120
40 to 44	541,897	27,018	49,913	70,090	102,429	128,647	146,327	212,378
45 to 54	967,090	51,130	51,463	74,107	107,043	140,679	156,779	252,797
45 to 49	518,077	26,719	51,463	73,555	106,197	139,350	155,398	250,847
50 to 54	449,013	24,411	50,975	74,209	108,779	142,213	159,296	258,997
55 to 62	538,661	30,701	45,989	69,028	103,599	134,097	153,986	242,685
55 to 59	369,224	20,980	47,789	71,152	104,984	137,530	155,398	252,797
60 to 62	169,437	9,721	42,479	64,802	99,097	126,617	145,038	217,406
63 & over	148,520	8,610	41,808	63,718	99,838	127,018	146,327	223,014
63 to 67	134,345	7,811	41,808	63,757	99,294	125,706	145,038	215,580
68 to 74	13,264	748	40,448	64,231	103,599	140,310	154,712	298,485
75 & over	n/r	51	10,870	51,463	77,194	127,042	156,734	254,873

State adjustments at mean
(State population must be >= 14,000 persons)

AK		MT	
AL	82.5	NC	91.8
AR	78.5	ND	
AZ	89.5	NE	83.4
CA	120.8	NH	100.0
CO	100.0	NJ	120.8
CT	118.8	NM	72.9
DC	114.9	NV	81.5
DE		NY	108.9
FL	85.0	OH	90.0
GA	94.3	OK	77.7
HI		OR	88.7
IA	83.9	PA	95.9
ID		RI	
IL	104.9	SC	77.3
IN	83.5	SD	
KS	83.0	TN	81.5
KY	74.5	TX	103.3
LA	79.9	US	100.0
MA	114.2	UT	92.9
MD	106.8	VA	107.4
ME		VT	
MI	95.0	WA	108.8
MN	96.2	WI	87.8
MO	84.0	WV	
MS	64.4	WY	

Notes:

All national estimates are produced from ACS non-allocated populations of 28,000 or more wage & salary employees who do not have income from self-employment, retirement, Social Security, Supplemental Security, or public assistance. The indicator 'n/r' signals that the estimates for the subgroup are not reliable. The indicator 'n/d' signals that no data is available within the subgroup. Care should be exercised in using and interpreting earnings values for small subgroups of the population.

Conforming age-earnings regression results

Level	α	β	γ	α error	β error	γ error	α t-stat	β t-stat	γ t-stat
10th	-21,423	2,578.41	-26.191	1,608	76.90	0.893	-13.3	33.5	-29.3
25th	-13,087	2,581.26	-24.494	1,244	62.57	0.749	-10.5	41.3	-32.7
Median	-31,352	3,971.17	-37.302	1,391	70.26	0.828	-22.5	56.5	-45.1
Mean	-63,312	5,818.38	-55.765	2,100	101.21	1.169	-30.2	57.5	-47.7
75th	-69,516	6,691.78	-65.080	2,276	112.18	1.307	-30.5	59.6	-49.8
90th	-126,559	10,471.41	-102.389	3,938	209.46	2.572	-32.1	50.0	-39.8

Cross-tabulation of earnings by age group

Ages	U.S. Pop	ACS Resp	10th	25th	Median	Mean	75th	90th
18 & over	4,182,068	219,587	36,582	48,691	64,843	77,190	90,575	124,319
18 to 62	4,046,012	211,480	36,582	48,537	64,802	76,907	90,268	123,511
18 to 24	34,289	1,635	16,179	27,175	40,448	41,733	52,835	64,802
18 to 19		n/d	n/d	n/d	n/d	n/d	n/d	n/d
20 to 24	34,289	1,635	16,179	27,175	40,448	41,733	52,835	64,802
25 to 34	1,124,786	55,040	31,356	41,808	53,305	60,959	72,519	96,158
25 to 29	458,178	22,208	28,142	38,332	49,124	53,139	63,705	82,879
30 to 34	666,608	32,832	36,024	45,169	57,638	66,334	78,390	104,073
35 to 44	1,238,750	62,812	40,355	50,763	67,339	79,257	93,029	125,569
35 to 39	642,568	32,196	39,436	50,169	64,843	76,045	90,131	121,343
40 to 44	596,182	30,616	40,448	51,799	69,772	82,719	96,063	131,455
45 to 54	1,036,892	56,296	41,170	53,305	72,519	86,794	101,119	138,950
45 to 49	556,662	29,434	41,170	52,835	72,048	85,756	99,838	135,875
50 to 54	480,230	26,862	40,763	53,871	73,164	87,997	101,119	141,566
55 to 62	611,295	35,697	41,273	55,222	73,276	86,695	100,339	135,499
55 to 59	422,285	24,632	41,440	54,604	73,164	87,511	101,119	138,056
60 to 62	189,010	11,065	40,922	55,580	74,107	84,873	98,419	131,684
63 & over	136,056	8,107	39,508	54,350	73,164	85,594	98,763	132,746
63 to 67	126,880	7,571	40,448	55,222	73,555	86,035	99,097	132,746
68 to 74	8,460	498	32,936	48,915	69,005	80,564	93,239	138,056
75 & over	n/r	38	15,930	23,363	52,260	66,957	77,194	111,231

State adjustments at mean
(State population must be >= 14,000 persons)

AK		MT	
AL	83.6	NC	87.2
AR	80.5	ND	
AZ	87.1	NE	87.4
CA	120.4	NH	100.0
CO	96.8	NJ	118.8
CT	112.6	NM	83.0
DC	129.2	NV	89.7
DE	100.0	NY	111.4
FL	85.6	OH	93.4
GA	90.4	OK	78.5
HI	96.3	OR	93.7
IA	93.1	PA	100.0
ID		RI	104.7
IL	104.3	SC	77.1
IN	88.2	SD	
KS	84.3	TN	83.2
KY	81.8	TX	98.4
LA	80.1	US	100.0
MA	113.4	UT	87.9
MD	113.6	VA	107.5
ME	81.3	VT	
MI	96.4	WA	105.7
MN	105.1	WI	95.4
MO	84.5	WV	81.9
MS	71.8	WY	

Notes:

All national estimates are produced from ACS non-allocated populations of 28,000 or more wage & salary employees who do not have income from self-employment, retirement, Social Security, Supplemental Security, or public assistance. The indicator 'n/r' signals that the estimates for the subgroup are not reliable. The indicator 'n/d' signals that no data is available within the subgroup. Care should be exercised in using and interpreting earnings values for small subgroups of the population.

Testing the Fit of a Regression Model

$FTE = \alpha + \beta \text{Age} + \gamma Age^2$ where FTE is full time earnings

Why would you use a quadratic functional form? Just look at the plots on previous page.

Why does the study divide individuals by Gender, Educational Attainment and by percentiles of FTE?
(Gauss Markov #4. standard deviation of ε, σ_ε, remains constant)

For Men in 90 percentile (in top 10%):

$FTE = \alpha + \beta \text{Age} + \gamma Age^2$ = −286,486 + 18,712(Age) – 162(Age^2)
(29.6) (-20) (t-ratios)

For Women in 90 percentile (in top 10%):

$FTE = \alpha + \beta \text{Age} + \gamma Age^2$ = −126,559 + 10,471(Age) – 102(Age^2)
(50) (-40) (t-ratios)

Do both β and γ test significant at 95% cl?

Yes, for both models: $|t - ratio| > 2$ **for both** Age **and** Age^2

So for Men, FTE peaks at what age? Max FTE where Age = -β/2γ = -18,712/2(-162) = 57.75

And for Women? Max FTE where Age = -β/2γ = -10,471/2(-102) = 51.33

For a 52 year old women, each additional year of age adds how much to FTE?

$\Delta FTE = (\beta + 2\gamma Age) = 10,471 + (2 * -102 * 52)$ = -137

For a 52 year old man, each additional year of age adds how much to FTE?

$\Delta FTE = (\beta + 2\gamma Age) = 18,712 + (2 * -162 * 52)$ = 1,864

When to consider Nonlinear, if regressors are expected to interact

- product of 2 variables, e.g., slope dummy (see below)
- Synergism (positive interaction): more than sum of parts
- Negative interaction: conflict, too many cooks spoil the broth
- Slope dummy: product of quantitative and dummy var
  - E.g. recall men earnings: $Wage = \beta_0 + \beta_1 Exp + \beta_2 Male + \varepsilon$
  - vs Male raises higher: $Wage = \beta_0 + \beta_1 Exp + \beta_2 ExpMale + \varepsilon$
  - use slope dummy: $ExpMale = Exp$ if man, $= 0$ if woman
  - Same as 2 equations: $Wage = \beta_0 + \beta_1 Exp + \varepsilon$ for women
    - and $Wage = \beta_0 + (\beta_1 + \beta_2) Exp + \varepsilon$ for men
  - Interactive because of the product: $ExpMale = Exp * Male$
  - Slope dummy: slope is β_1 for women, $(\beta_1 + \beta_2)$ for men

When to consider Nonlinear, if regressors are expected to interact (continued)

- Dummy variable interaction Ex:
  $Salary = \beta_0 + \beta_1 MBA + \beta_2 Eng + \beta_3 Both + \varepsilon$
- Where MBA and Engr are college degree dummies and
  $Both = MBA * Eng$ which equals 1 if have both degrees
  - E.g.: $Salary = 30 + 25 MBA + 15 Eng + 20 Both + \varepsilon$

Means that MBAs earn \$25K more, engineers \$15K more, but with both degrees, earn $(25 + 15 + 20) =$ \$60K more But if $b_3 < 0$, negative interaction, whole $<$ sum of parts

Multiplicative Functional Form

- $Y = aX^bZ^c$ nonlinear, so is linear regression hopeless? Not hopeless as you can easily transform to linear by taking natural logs. Recall the rules for logs:

$$\ln xz = \ln x + \ln z \text{ and } \ln(x^a) = a \ln x$$

- So transform $Y = aX^bZ^c$ into $\ln Y = \ln a + b \ln X + c \ln Z$
  - Note: can't compare R2's with linear because different dep var!
- For demand, regress logarithms so that coefficients are elasticities
  $LQCar = 3.0 - 2LPCar - 0.5LPGas + +0.4LPImprt + 4LInc$
  - Price elasticity $EP = -2$, +5% price cuts demand 2(5%) = 10%
  - Elasticity for gas price, import price, income = -0.5, +0.4, +4

Other Functional Forms

- Ratios: e.g., financial ratios, percent Latino
- %Change or Difference: especially time series
- Reciprocal or square root: alternative to logs

When to Consider Nonlinear

- Indicated by theory or logic (e.g., log, quadratic)
- Variable collected over wide enough range of values
- Enough degrees of freedom: n much larger than k
- Easier to interpret results (e.g., demand elasticities)
- error analysis (residual plot) or variable plot