GOV 2000 Section 8: Diagnosing and Fixing Regression Problems

Konstantin Kashin[1]
Harvard University

October 24, 2012

[1]These notes and accompanying code draw on the notes from Molly Roberts, Maya Sen, Iain Osgood, Brandon Stewart, and TF's from previous years

Outline

Administrative Details

- Midterm returned
- Problem Set 5 returned; corrections due next Tuesday
- Problem Set 7 due next Tuesday

Outline

What have we covered?

- Summarizing and describing data: both univariate and multivariate populations
- Sampling as source of randomness and uncertainty
  - Probability / random variables
  - Sample statistics and sampling distributions
- Revisit regression in context of sampling
  - Standard errors, hypothesis testing, and confidence intervals for estimated regression coefficients
- But wait! There are many assumptions that go into regression...
- And there's missing data...

Where are we going?

Causal inference is a missing data problem!

Outline

What are key regression assumptions?

- Random sampling
- Constant variance (homoskedasticity)
- Normality
- Linear conditional expectation function

Credit Card Expenditures Data

We will be working with `ccarddata.csv`.

- Outcome variable: credit card expenditure
- Covariates:
  - Age
  - Household income (monthly in thousands of dollars)
  - Dummy for home ownership

OLS

```
lm.cc <- lm(ccexpend ~ income + homeowner + age, data=
    cc)
```

OLS

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3972	163.9851	0.00	0.9981
income	79.8358	23.6724	3.37	0.0012
homeowner	32.0877	84.7241	0.38	0.7061
age	-0.7769	5.5128	-0.14	0.8883

Violations of Constant Variance Assumption

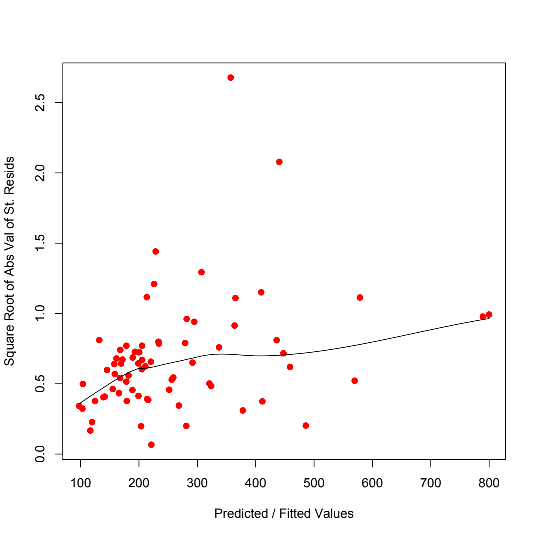
Why is heteroskedasticity an issue?

- Estimated variances / standard errors are biased
- OLS is no longer efficient (BLUE)
- Hypothesis testing and confidence intervals are off

Good news: problem is usually not that bad (depends on severity of heteroskedasticity) and point estimates of regression coefficients still unbiased!

Diagnosing Heteroskedasticity

We can use a scale-location plot to diagnose heteroskedasticity:



Diagnosing Heteroskedasticity

In R, we can construct a scale-location plot as follows:

```
plot(lm.cc, 3)
```

Or manually as:

```
scatter.smooth(fitted(lm.cc), sqrt(abs(rstudent(lm.cc)
    )), col="red")
```

Fixing Heteroskedasticity

- **Treat it as a nuisance:** use heteroskedasticity-consistent standard errors (i.e. Huber-White)
- **Model it:** use Weighted Least Squares (WLS)
- **Treat it as model diagnostic tool**: change entire model

Homoskedastic Variance-Covariance Matrix

How can we characterize the variance of the error terms under homoskedasticity?

$$V[\boldsymbol{\epsilon}] = \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{pmatrix}$$

We then estimate σ^2 with $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n} \hat{\epsilon}^2}{n-k-1}$

Heteroskedastic Variance-Covariance Matrix

How can we characterize the variance of the error terms under heteroskedasticity?

$$V[\boldsymbol{\epsilon}] = \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_3^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}$$

How do we estimate the σs?

Huber-White Variance-Covariance Matrix

The Huber-White variance-covariance matrix is a **consistent estimate** of the heteroskedastic $\boldsymbol{\Sigma}$:

$$\hat{V}[\boldsymbol{\epsilon}] = \hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \widehat{\epsilon_1^2} & 0 & 0 & \cdots & 0 \\ 0 & \widehat{\epsilon_2^2} & 0 & \cdots & 0 \\ 0 & 0 & \widehat{\epsilon_3^2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \widehat{\epsilon_n^2} \end{pmatrix}$$

Fixing Heteroskedasticity: Huber-White SEs

In R, we can calculate Huber-White SEs as:

```
library(car)

# returns variance-covariance matrix:
hccm(lm.cc, type="hc0")

# returns standard errors:
sqrt(diag(hccm(lm.cc, type = "hc0")))
```

Fixing Heteroskedasticity: Small Sample Correction

A potential problem with Huber-White SEs is that it requires a **larger sample size**. Thus, we often use a **small-sample correction** to obtain more conservative estimates:

$$\hat{\mathbb{V}}[\boldsymbol{\epsilon}] = \hat{\boldsymbol{\Sigma}} = \frac{n}{n-k-1}\begin{pmatrix} \widehat{\epsilon_1^2} & 0 & 0 & \cdots & 0 \\ 0 & \widehat{\epsilon_2^2} & 0 & \cdots & 0 \\ 0 & 0 & \widehat{\epsilon_3^2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \widehat{\epsilon_n^2} \end{pmatrix}$$

This is often known as HC1 and is used in many publications (`,robust` option in Stata).

Fixing Heteroskedasticity: Small Sample Correction

In R, we can calculate Huber-White SEs with different small-sample corrections as:

```
hccm(lm.cc, type="hc1")
hccm(lm.cc, type="hc2")
hccm(lm.cc, type="hc3")
```

Heteroskedastic Variance-Covariance Matrix

We can characterize the variance of the error terms under heteroskedasticity as weighted homoskedastic matrix:

$$V[\boldsymbol{\epsilon}] = \boldsymbol{\Sigma} = \begin{pmatrix} a_1^2\sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & a_2^2\sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & a_3^2\sigma^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_n^2\sigma^2 \end{pmatrix} = \sigma^2 \begin{pmatrix} a_1^2 & 0 & 0 & \cdots & 0 \\ 0 & a_2^2 & 0 & \cdots & 0 \\ 0 & 0 & a_3^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_n^2 \end{pmatrix}$$

If we knew weights a_i, we could reweight data using $\frac{1}{a_i}$. Problem is we don't know the weights exactly...

Fixing Heteroskedasticity: Weighted Least Squares (WLS)

An alternative to heteroskedasticity-consistent standard errors is using WLS whereby we weight observations that we believe have small error variance higher.

- Efficient **if** we correctly specify weights
- We may have good guess as to what weights are
- Unbiased for βs and consistent for $V(\boldsymbol{\beta})$ even if we get weights wrong

If we, for example, believe that error variance is inversely proportional to income:

```
lm.cc.wt <- lm(lm.cc$call, weights=1/income, data=cc)
```

Violations of Normality Assumption

Why is non-normality an issue?

In small samples:

- $\hat{\beta}$ will not have normal sampling distribution
- Test statistics will not have t distributions
- Since SEs are off, we have incorrect probability of Type I error in testing and incorrect coverage of confidence intervals

Good news: **in large samples**, Central Limit Theorem makes these problems go away!

Diagnosing Non-Normality

- **Density plots of errors:** studentized residuals should have t distribution with $n - k - 2$ degrees of freedom
- **Formal tests**
- **Quantile-quantile (Q-Q) plots**

Diagnosing Non-Normality: Q-Q Plots

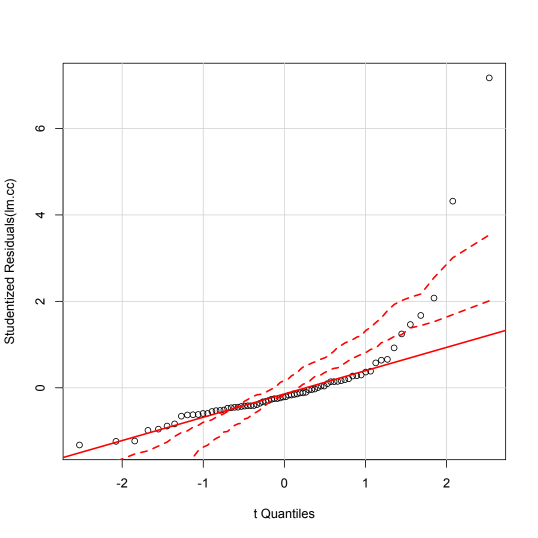
- **Generally:** we compare quantiles of empirical distribution with quantiles of theoretical distribution
- **Specifically:** we compare quantiles of studentized residuals to quantiles of t distribution

Diagnosing Non-Normality: Q-Q Plots

In R:

```
library(car)
qqPlot(lm.cc)

plot(lm.cc,2)
```

Diagnosing Non-Normality: Q-Q Plots



Violations of Linearity Assumption

Why is non-linearity an issue?

- Bias in estimated regression function (for population CEF, not for best linear approximation to CEF)
- Leads to bad inferences and poor prediction

Diagnosing Non-Linearity: GAM Plots

Fitting a generalized additive model (GAM) can reveal non-linearities:

```
library(mgcv)
gam.cc <- gam(ccexpend ~ s(income) + s(age) +
    homeowner, data=cc)
```

- Using `s()` around the variables allows GAM to choose smooth functional form
- Algorithm minimizes deviations from surface without fitting data too closely (bias-variance tradeoff)

Diagnosing Non-Linearity: GAM Plots

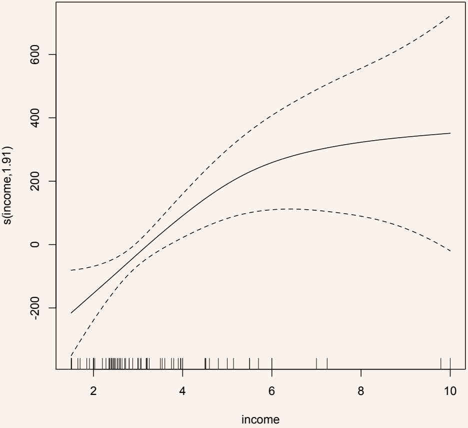
```
> summary(gam.cc)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   252.06      45.78   5.506 6.26e-07 ***
homeowner      27.94      83.10   0.336    0.738
---

Approximate significance of smooth terms:
            edf Ref.df     F p-value
s(income) 1.912   2.38 6.151 0.00229 **
s(age)    1.000   1.00 0.171 0.68057
---

R-sq.(adj) =  0.199   Deviance explained = 24.3%
GCV score =  86930  Scale est. = 81000     n = 72
```

Diagnosing Non-Linearity: GAM Plots

- Equilalent degrees of freedom (edf): how many variables are needed to define smooth regression surface
  - edf= k: smooth surface is linear
  - edf> k: smooth surface deviates from linear (and requires additional variables)
  - Our example: edf for `income` is ≈ 2, indicating that we should consider squared term
- F-value / p-value: probability that variable would have at least this extreme an effect under null hypothesis that there is no relationship
- Generalized cross-validation (GCV) score = predictive (out-of-sample) performance of smooth regression surface

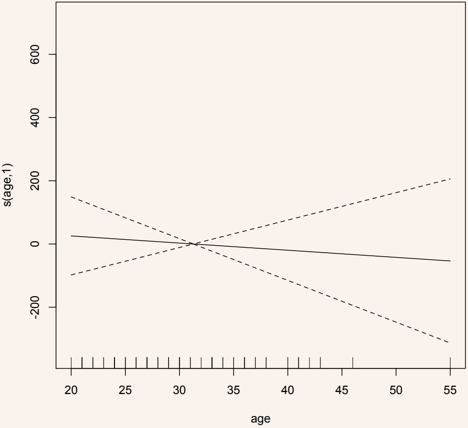
Diagnosing Non-Linearity: GAM Plots

We can also examine partial relationships between explanatory variables and the outcome graphically!



Diagnosing Non-Linearity: GAM Plots

We can also examine partial relationships between explanatory variables and the outcome graphically!



Fixing Non-Linearity

- Transform outcome variable (i.e. using log or square root)
- Transform explanatory variables (i.e. using log or square root) or add higher order terms
- Use semi-parametric or nonparametric models (i.e. GAMs)

Outline

Missingness Mechanism

How was missingness generated?

We can characterize the missingness mechanism as:

- Missing completely at random (MCAR): missingness unrelated to variables in data
- Missing at random (MAR): missingness related to observed data
- Not missing at random (NMAR): missingness related to unobserved data

Missingness in Our Data

Let's work with `ccarddata_missing.csv`, which now has missing values of credit card expenditures (outcome) for some observations.

How can we characterize this missingness?

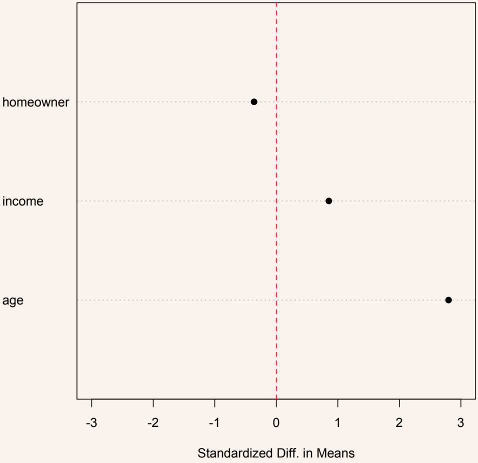
	$\bar{X}_{\text{missing}}$	$\bar{X}_{\text{non-missing}}$	$\bar{X}_{\text{missing}} - \bar{X}_{\text{non-missing}}$	t-stat
`age`	33.68	29.13	4.54	2.80
`income`	3.62	3.27	0.34	0.85
`homeowner`	0.35	0.39	-0.04	-0.36

Missingness in Our Data

We can also look at missingness mechanism graphically:

```
### create indicator for missingness
cc.missing$missing <- 0
cc.missing[is.na(cc.missing$ccexpend),]$missing <-1

### create vector of t-stats and plot
t.stats <- c(t.test(cc.missing[cc.missing$missing==1,"
    age"],cc.missing[cc.missing$missing==0,"age"])$
    statistic
,t.test(cc.missing[cc.missing$missing==1,"income"],cc.
    missing[cc.missing$missing==0,"income"])$statistic
,t.test(cc.missing[cc.missing$missing==1,"homeowner"],
    cc.missing[cc.missing$missing==0,"homeowner"])$
    statistic)

dotchart(t.stats, labels=c("age","income","homeowner")
    , xlim=c(-3,3), xlab="Standardized Diff. in Means"
    ,pch=19)
abline(v=0, col="red", lty=2)
```

Missingness in Our Data



Missingness in Our Data

The first 6 observations in our dataset are:

`ccexpend`	`age`	`income`	`homeowner`
124.98	38	4.52	1
	33	2.42	0
15.00	34	4.50	1
	31	2.54	0
546.50	32	9.79	1
92.00	23	2.50	0

Missing values shaded in red.

Dealing with Missingness

A few common ways to deal with missingness:

- Complete case analysis
- Mean imputation
- Regression imputation
- Multiple imputation

Complete Case Analysis

ccexpend	age	income	homeowner
124.98	38	4.52	1
	33	2.42	0
15.00	34	4.50	1
	31	2.54	0
546.50	32	9.79	1
92.00	23	2.50	0

In R:

```
# R automatically row-deletes observations with
    missing data
lm(ccexpend ~ income + homeowner + age, data=cc.
    missing)
```

Mean Imputation

ccexpend	age	income	homeowner
124.98	38	4.52	1
$\bar{y}$	33	2.42	0
15.00	34	4.50	1
$\bar{y}$	31	2.54	0
546.50	32	9.79	1
92.00	23	2.50	0

In R, mean of outcome is:

```
mean(cc.missing$ccexpend, na.rm=TRUE)
# mean is 209.4542
```

Mean Imputation

ccexpend	age	income	homeowner
124.98	38	4.52	1
209.45	33	2.42	0
15.00	34	4.50	1
209.45	31	2.54	0
546.50	32	9.79	1
92.00	23	2.50	0

Regression Imputation

ccexpend	age	income	homeowner
124.98	38	4.52	1
$\hat{y}_2$	33	2.42	0
15.00	34	4.50	1
$\hat{y}_4$	31	2.54	0
546.50	32	9.79	1
92.00	23	2.50	0

In R, we can predict missing values:

```
lm.cc.missing <- lm(ccexpend ~ income + homeowner +
    age, data=cc.missing)
missing.df<- cc.missing[is.na(cc.missing$ccexpend),c("
    income","homeowner","age")]
predict(lm.cc.missing,missing.df)
```

Regression Imputation

ccexpend	age	income	homeowner
124.98	38	4.52	1
94.41	33	2.42	0
15.00	34	4.50	1
109.15	31	2.54	0
546.50	32	9.79	1
92.00	23	2.50	0

Multiple Imputation

What limitations of previous imputation methods does multiple imputation address?