36-401 Class Notes and Exercise for 9/18/01

Simple Linear Regression

- *Model*: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2) \quad \epsilon_i \perp\!\!\!\perp \epsilon_j$
  Design or covariate space
  Estimate/Prediction: $\hat{Y}_i = b_0 + b_1 X_i \quad \hat{Y}^* = b_0 + b_1 X^*$
  Residual error: $e_i = Y_i - \hat{Y}_i$

- *Estimation method*: "Least Squares"

$$\mathrm{RSS} = \sum_{i=1}^{n} e_i^2 = \sum \left(Y_i - \hat{Y}_i\right)^2 = \sum \left(Y_i - b_0 - b_1 X_i\right)^2$$

  Minimization of RSS gives

$$b_1 = \frac{\sum \left[(X_i - \bar{X})(Y_i - \bar{Y})\right]}{\sum (X_i - \bar{X})^2} \quad b_0 = \bar{Y} - b_1 \bar{X}$$

- *Residual Variance* σ^2 estimated by $s_e^2 = \frac{\mathrm{RSS}}{n-K}$ $\quad s_e$ is "*residual standard error*"

- *Coefficient of Determination*: $\mathrm{R}^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}$
  Note: RSS+ESS=TSS if $\hat{Y}$ is based on LS estimates of b_0 and b_1.

- *Sample Covariance* of X and Y

$$s_{xy} = \frac{\sum \left[(X_i - \bar{X})(Y_i - \bar{Y})\right]}{n-1}$$

  Sample Correlation of X and Y

$$r = \frac{s_{xy}}{\sqrt{s_x^2}\sqrt{s_y^2}}$$

- *Sampling Distributions and Standard Errors*
  Assumptions on ϵ: normal, independent, same σ^2
  Estimated standard error of b_1:

$$\mathrm{SE}_{b_1} = s_e \sqrt{\frac{1}{\mathrm{TSS}_x}}$$

  Estimated standard error of b_0:

$$\mathrm{SE}_{b_0} = s_e \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\mathrm{TSS}_x}}$$

  Estimated standard error of prediction of $Y^* = b_0 + b_1 X^*$:

$$\mathrm{SE}_{Y^*} = s_e \sqrt{\frac{1}{n} + \frac{(X^* - \bar{X})^2}{\mathrm{TSS}_x} + 1}$$

  (drop +1 for estimated standard error of mean of Y for a given X)

More on Sampling Distributions

$$Y_i \sim \text{N}(\beta_0 + \beta_1 X_i,\ \sigma^2) \quad Y_i \perp\!\!\!\perp Y_j$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \sim \text{MVN}_n \left(\begin{bmatrix} \beta_0 + X_1\beta_1 \\ \beta_0 + X_2\beta_1 \\ \vdots \\ \beta_0 + X_n\beta_1 \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} \right) \equiv \text{MVN}_n \left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n \right)$$

—aside—

$$b_1 = \frac{\sum\left[(X_i - \bar{X})(Y_i - \bar{Y})\right]}{\sum(X_i - \bar{X})^2} = \frac{1}{\sum(X_i - \bar{X})^2}\left\{\sum\left[(X_i - \bar{X})Y_i\right] - \sum\left[(X_i - \bar{X})\bar{Y}\right]\right\}$$

$$= \frac{\sum\left[(X_i - \bar{X})Y_i\right]}{\sum(X_i - \bar{X})^2}$$

$$\text{Var}\left(\frac{(X_1 - \bar{X})}{\sum(X_i - \bar{X})^2} Y_1\right) = \sigma^2 \frac{(X_1 - \bar{X})^2}{\left[\sum(X_i - \bar{X})^2\right]^2}$$

$$\Rightarrow \text{Var}(b_1) = \sigma^2 \frac{\sum(X_i - \bar{X})^2}{\left[\sum(X_i - \bar{X})^2\right]^2} = \frac{\sigma^2}{\text{TSS}_X}$$

—end aside—

$$b_1 \sim \text{N}\left(\beta_1, \frac{\sigma^2}{TSS_X}\right) \Rightarrow \frac{b_1 - \beta_1}{\sigma/\sqrt{TSS_X}} \sim \text{N}\left(0, 1\right)$$

$$\frac{b_1 - \beta_1}{s_e/\sqrt{TSS_X}} \sim \text{t}_{n-2}$$

What We Get When Assumptions Are Met

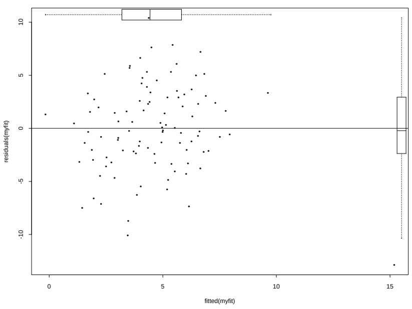
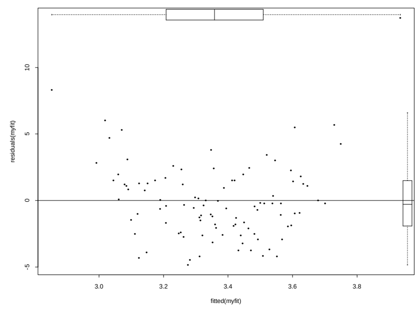
- Estimates of coefficients (b_0 and b_1) are unbiased
- Estimate of residual variance (s_e^2) is unbiased
- Above estimates are minimum variance (most efficient)
- Standard errors of the coefficients are unbiased
- t and F tests are valid

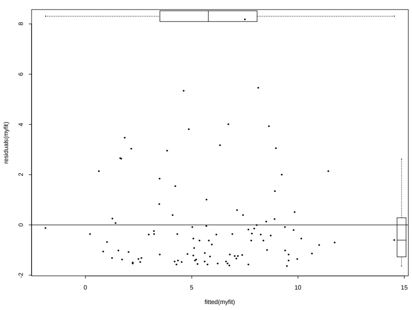
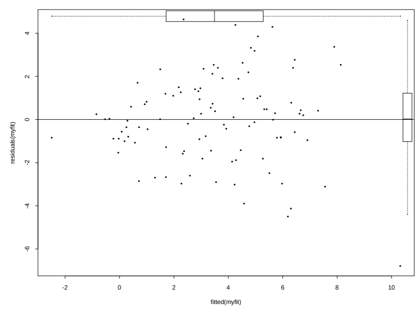
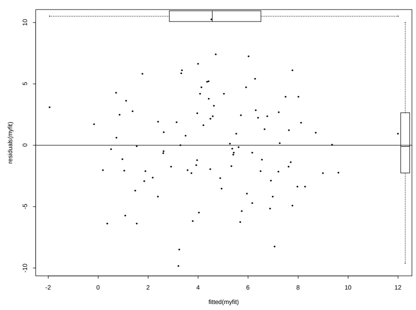
Regression Problems, Residual Analysis

- *Common Regression Problems*
  - Omitted Variables
  - Influential Observations
  - Nonlinear Relationships
  - Nonnormal Residuals
  - Nonconstant Residual Variance
  - Correlation Among Residuals

- *Diagnistic Plots*
  - Univariate EDA of X, Y: histograms, boxplots, quantile-normal plots, etc.
  - Y_i vs X_i with the regression line ($\hat{Y}_i$ vs X_i) overlaid:
    `plot(X,Y); myfit.lm <- lm(Y ~ X); abline(myfit)`
  - e_i vs X_i: `plot(residuals(myfit.lm), X); abline(h=0)`
  - e_i vs $\hat{Y}_i$: `plot(residuals(myfit.lm), fitted(myfit.lm)); abline(h=0)`
  - EDA of the residuals: histograms, boxplots, quantile-normal plots, etc.

- *Residual Analysis*
  - A good residual plot is an amorphous, football shaped blob centered on the line $e = 0$.
  - All of the following should draw your attention: outliers, curvature, skewing or clumping (in X or Y!), heteroscedasticity.

Testing and Confidence Intervals for Simple Regression

```
> homedat _ read.table("homedat.dat",header=T)
> names(homedat)
[1] "PRICE" "SQFT"  "AGE"   "FEATS" "NE"    "CUST"  "COR"   "TAX"
> myfit.lm _ lm(PRICE ~ SQFT,data=homedat)
> attach(homedat)
> par(mfrow=c(2,3))
> qqenv(SQFT)
> qqenv(PRICE)
> scatbox(SQFT,PRICE)                          ; abline(myfit.lm)
> scatbox(SQFT,residuals(myfit.lm))            ; abline(h=0)
> scatbox(fitted(myfit.lm),residuals(myfit.lm)) ; abline(h=0)
> qqenv(residuals(myfit.lm))
> # NOW MAYBE TRY better.lm _ lm(log(PRICE) ~ sqrt(SQFT),data=homedat)

> summary(myfit.lm)
[ ... ]

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept) 47.8193 62.8548     0.7608  0.4483
       SQFT  0.6137  0.0362    16.9306  0.0000

Residual standard error: 204.5 on 115 degrees of freedom
Multiple R-Squared: 0.7137
F-statistic: 286.6 on 1 and 115 degrees of freedom, the p-value is 0

[ ... ]
```

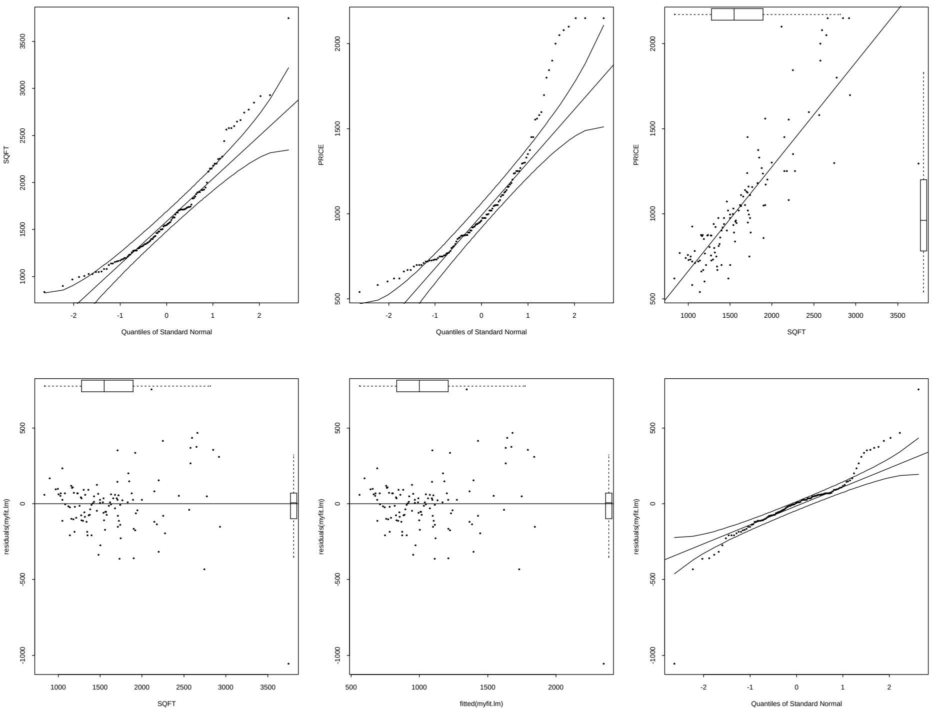
- t-test and meaning of p-value

$$\frac{b - \beta_h}{\text{estSE}_b} \sim t_{n-K} \quad \text{p-value} = 2(1-\text{pt}(\text{abs}(t),\ n-K))$$

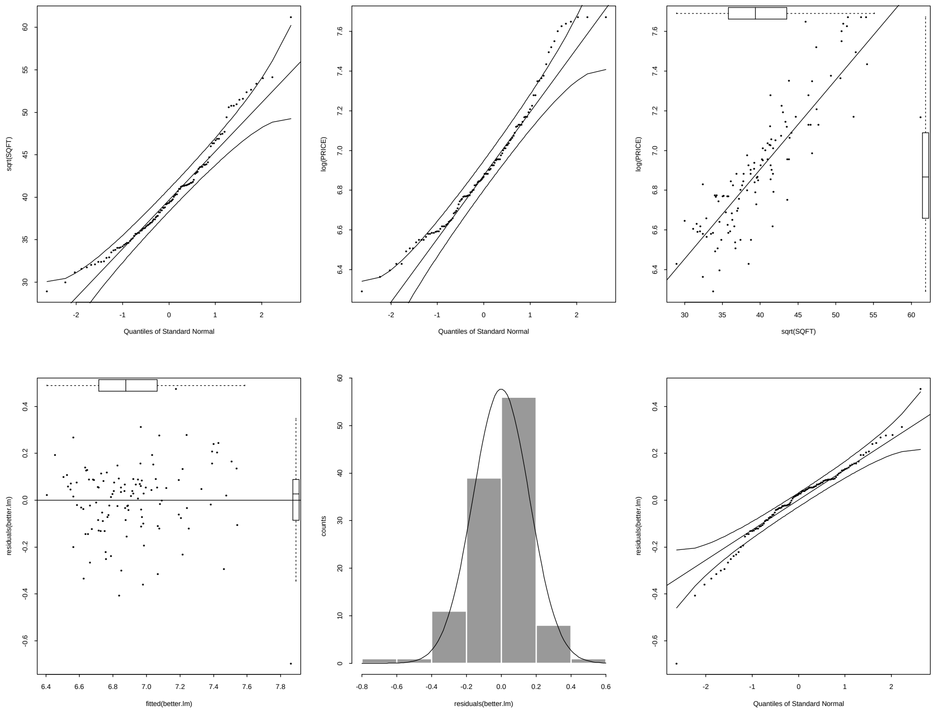
- Confidence intervals for coefficients
  Need, e.g., $t_{(0.05,\text{df})} = \text{qt}(1 - 0.05/2, \text{df})$
  CI is $b_1 \pm t_{(0.05,\text{df})}\text{estSE}_{b_1}$

- Prediction interval for $Y \mid X = X_i$
  CI is $\mu_i \pm t_{(0.05,\text{df})}\text{estSE}_{Y_i^*}$

- Confidence interval for $E(Y \mid X = X_i)$
  for CI drop “1” from formula for $\text{estSE}_{Y_i^*}$



Some diagnostic plot for $PRICE = \beta_0 + \beta_1 SQFT + \epsilon$.



Some diagnostic plot for $\log(PRICE) = \beta_0 + \beta_1 \sqrt{SQFT} + \epsilon$.

ANOVA Table

- ESS (Explained SS), EMS (Explained MS), RSS (Residual SS), RMS (Residual MS), df…

Source	DF	SS	MS	F	p-value
Model	1	ESS	EMS = $\frac{ESS}{1}$	$\frac{EMS}{RMS}$	$P[F > \frac{EMS}{RMS}]$
Resid	$n-2$	RSS	RMS = $\frac{RSS}{n-2}$		
Total	$n-1$	TSS_Y	TMS = $\frac{TSS_Y}{n-1}$		

```
> anova(myfit.lm)
Analysis of Variance Table

Response: PRICE

Terms added sequentially (first to last)
           Df Sum of Sq  Mean Sq  F Value Pr(F)
     SQFT   1  11981915 11981915 286.6468     0
Residuals 115   4807032    41800
```

- "Omnibus" F-test
  F-statistic: 286.6468 on 1 and 115 degrees of freedom, the p-value is essentially zero

Regression Through the Origin

```
> zerofit.lm _ lm(PRICE ~ SQFT - 1, data=homedat)
> summary(zerofit.lm)

Residuals:
   Min     1Q Median    3Q   Max
 -1105 -85.8  14.45 84.55 745.8

Coefficients:
       Value Std. Error t value Pr(>|t|)
SQFT  0.6400  0.0109    58.8206  0.0000

Residual standard error: 204.1 on 116 degrees of freedom
Multiple R-Squared: 0.9676
F-statistic: 3460 on 1 and 116 degrees of freedom, the p-value is 0
```

- Misleading R^2 and F

Breakout Session

Exercise 1

The prevalence of respiratory symptoms was recorded for 9 groups of subjects exposed to differing levels of dust in their work environment. Dust exposure was measured as particles/ft^3/year scaled by 10^6. The direct outcome variable is "relative risk", the ratio of symptom prevalence at a given exposure level to symptom prevalence in the absence of workplace dust.

```
> dust_data.frame(exposure=c(75,100,150,350,600,900,1300,1650,2250),
                RR=c(1.10,1.05,0.97,1.90,1.83,2.45,3.70,3.52,4.16))
> summary( lm(RR ~ exposure, dust) )
Coefficients:  Value Std. Error t value Pr(>|t|)
(Intercept)   1.0360 0.1688     6.1358  0.0005
   exposure   0.0015 0.0002     9.9927  0.0000

Residual standard error: 0.3363 on 7 degrees of freedom
Multiple R-Squared: 0.9345
F-statistic: 99.85 on 1 and 7 degrees of freedom, the p-value is 2.15e-05
```

Is the intercept different from 0? What is your guess at the "correct" value of the intercept just from reading the experimental protocol? Why? What is the t-value to test if the intercept is different from that value? Do you think this represents a "significant" t-value?

Exercise 2

```
> summary( lm(RR-1 ~ exposure, dust) )
Residuals:
     Min    1Q   Median      3Q    Max
 -0.3406 -0.14 -0.05667 0.02818 0.6623

Coefficients:  Value Std. Error t value Pr(>|t|)
(Intercept)   0.0360 0.1688     0.2132  0.8373
   exposure   0.0015 0.0002     9.9927  0.0000

Residual standard error: 0.3363 on 7 degrees of freedom
Multiple R-Squared: 0.9345
F-statistic: 99.85 on 1 and 7 degrees of freedom, the p-value is 2.15e-05
```

(Note that -1 on the left side of the tilde in the regression formula just subtracts one from the Y variable.) Is the intercept significantly different from zero in the second regression run? What is a 95% CI (confidence interval) for b_0?

Exercise 3

```
> summary( lm(RR-1 ~ exposure-1, dust) )
Call: lm(formula = RR - 1 ~ exposure - 1, data = dust)
Residuals:
     Min      1Q   Median      3Q    Max
 -0.3598 -0.1086 -0.06121 0.04207 0.6663

Coefficients:
           Value Std. Error t value Pr(>|t|)
exposure  0.0016  0.0001    16.2955  0.0000

Residual standard error: 0.3156 on 8 degrees of freedom
Multiple R-Squared: 0.9708
F-statistic: 265.5 on 1 and 8 degrees of freedom, the p-value is 2.024e-07
```

Is the third regression better than the second? What do we need to do to see if the model is adequate? How do you interpret the model substantively, i.e. how could you explain this to a statistically naive researcher in Environmental Studies?