17. Simple Linear Regression Diagnostics

Readings: Kleinbaum, Kupper, Nizam, and Rosenberg (KKNR): Ch. 14

SAS: PROC REG

Homework: Homework 6 due by 11:59 pm on October 24
Homework 7 due by 11:59 pm on October 31
Homework 8 due by 11:59 pm on November 7

Overview

A) Re/Preview of Topics
B) Linear Regression Assumptions Revisited
C) Residuals
D) Diagnostic Plots
E) Transformations to Remove Heteroscedasticity
F) Regression Diagnostics Recap

A. Review (Lectures 15-16)/ Current (Lecture 17)/ Preview (Lecture 18)

Lecture 16-17:

- Fit a line to data: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ and derived $\hat{\beta}_0, \hat{\beta}_1, Var(\hat{\beta}_0), Var(\hat{\beta}_1)$ by minimizing $SS_{Error} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$
- Partitioning out the Regression and Residual Components: $SS_{Total} = SS_{Model} + SS_{Error}$
- ANOVA vs parameter table for simple linear regression
- R^2: Proportion of the variance of Y that can be explained by the variable X
- Prediction Intervals (larger variance) and Confidence Intervals (smaller variance)

Lecture 17:

- Diagnostics to evaluate linear regression assumptions
- Data transformations to address violations of the assumptions

Lecture 18:

- Simple Linear Regression Examples (one covariate)
- Motivation for Multiple Linear Regression (more than one covariate)

B. Linear Regression Assumptions Revisited

Existence: For any fixed value of the variable *X*, *Y* is a random variable with a certain probability distribution having finite mean and variance.

Independence: The *Y*-values are statistically independent of one another.

Linearity: The mean value of *Y*, $\mu_{Y|X}$, is (approximately) a straight-line function of *X*.

Homoscedasticity: The variance of *Y* is the same for any *X*. That is,

$$\sigma^2_{Y|X} = \sigma^2_{Y|X=1} = \sigma^2_{Y|X=2} = \ldots = \sigma^2_{Y|X=x}$$

Normal Distribution: For any fixed value of *X*, *Y* has a normal distribution.
Note this assumption does ***not*** state that *Y* is normally distributed, but that *Y*|*X* is normally distributed.

Recall the normality assumption

- Not required to obtain estimate of the regression coefficients, β's
- Needed to perform statistical tests (*t*- or *F*-tests depend on normality assumption)
- Obtain confidence intervals (rely on *t*- or *F*-distributions, which assume normality)
- Estimates are asymptotically normal (i.e. assume normal when sample size is large)

Regression Diagnostics

1. Regression diagnostics are tools that can be used to assess the linearity, homoscedasticity, and normality assumptions of linear regression. (Used more often)

2. Regression diagnostics are also used to help identify outliers and influential points in a regression model. (Used less often)

With linear models, the assumptions of linearity, homoscedasticity, and normality are so intertwined that they often are met or violated as a set.

On the other hand, actions taken to correct violations of one assumption may result in violations of another assumption (e.g., transformations to stabilize the variance can lead to non-linearity).

Prior to fitting a regression model, simple descriptive statistics should be evaluated to look for data errors, potential outliers, and other potential violations of assumptions:

- Univariate descriptive statistics (mean, SD, min, max; frequency tables; histograms).
- Bivariate descriptive statistics (correlations/scatterplots).

C. Residuals

In regression analysis we assume that the *unobserved* error terms (ε_i):

- Are independent (uncorrelated).
- Have a mean of zero.
- Have a common variance $\sigma^2_{Y|X}$.
- Follow a normal distribution (required for performing parametric tests of significance and for calculating confidence intervals).

Recall that each of our *observed* residuals ($\hat{e}_i = Y_i - \hat{Y}_i$) are estimates of the *unobserved* error terms (ε_i).

Note that the $\hat{e}_i$ are not independent random variables (since they must sum to zero).

- In general, if the number of residuals (*n*) is large relative to the number of independent variables (*p*), the dependency effect can, for all practical purposes, be ignored in any analysis of the residuals.

We can examine the observed residuals (or functions of observed residuals) to assess the appropriateness of the assumptions listed above.

Types of residuals

Observed residual: The difference between the observed and predicted values $\hat{e}_i = Y_i - \hat{Y}_i$.

- The observed residuals have a mean of 0 and variance

$$S_e^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}\hat{e}_i^2$$

which is the residual mean square (MSE).

- The magnitude of the observed residuals depends on the scaling of *Y*, and thus different methods of standardizing residuals have been developed:
    - Standardized residuals
    - Studentized residuals
    - Press residuals
    - Jackknife residuals (**Most Preferred**)

<u>Standardized residual (Semi-studentized residual)</u>: observed residual divided by $\hat{\sigma}_{Y|X}$.

$$z_i = \frac{\hat{e}_i}{\hat{\sigma}_{Y|X}} = \frac{\hat{e}_i}{\sqrt{MSE}}$$

- Standardized residuals have a mean of zero and a variance of 1 (i.e., the standard normal distribution)

__Studentized residual__: observed residual divided by the standard deviation of the *i*th residual $[Var(\hat{e}_i) = MSE \times (1 - h_i)]$:

$$r_i = \frac{\hat{e}_i}{\sqrt{MSE \times (1-h_i)}} = \frac{z_i}{\sqrt{(1-h_i)}} = \frac{(\text{standardized residual})_i}{\sqrt{(1-h_i)}}$$

Matrix Note:
$\hat{Y} = HY$

where the quantity h_i, known as the **leverage**, is the *i*th element on the diagonal of the hat matrix.

- The hat matrix, *H*, is calculated as $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ (more in Lecture 20).
- The leverage is a measure of the importance of the *i*th observation in determining the model fit (more in Lecture 26).

- Studentized residuals have a mean near 0 (but not exactly 0), and variance

$$\frac{1}{n-p-1}\sum_{i=1}^{n} r_i^2$$

  that is slightly larger than 1.

- In large data sets, the standardized and Studentized residuals should not differ dramatically.

- The Studentized residuals follow <u>approximately</u> a *t* distribution with *n*-*p*-1 degrees of freedom (assuming the assumptions about the errors are satisfied).

***Deleted residual* (*Press residual*):** standardized residual with the current observation deleted from the calculation of the β’s (and thus from the calculation of the MSE). The *deleted residual* is

$$\hat{d}_i = Y_i - \hat{Y}_{i(-i)}$$

where $\hat{Y}_{i(-i)}$ is the predicted value for the i^{th} observation from a model fit *without* it (i.e., deleted from that model for estimating the β’s and predicted using the resulting estimates).

To avoid fitting *n* different regression models to calculate the deleted residual, we can instead use the following formula

$$\hat{d}_i = \frac{\hat{e}_i}{1 - h_{ii}}$$

where h_{ii} is the diagonal element from the **H** matrix.

***Jackknife residual (Studentized deleted residual, R-student, Studentized Press, Externally Studentized)*:** the Studentized residual with the current observation deleted:

$$r_{(-i)} = r_i \sqrt{\frac{MSE}{MSE_{(-i)}}} = \frac{\hat{e}_i}{\sqrt{MSE_{(-i)}(1-h_i)}} = r_i \sqrt{\frac{(n-p-1)-1}{(n-p-1)-r_i^2}}$$

where $MSE_{(-i)}$ is the residual variance (MSE) computed with the *i*th observation deleted.

- Jackknife residuals have a mean near 0 and a variance

  $$\frac{1}{(n-p-1)-1} \sum_{i=1}^{n} r_{(-i)}^2$$

  that is slightly greater than 1.

- Jackknife residuals exactly follow a *t* distribution with (*n*-*p*-1)-1 degrees of freedom.

Examining Residuals

If the standard regression assumptions are satisfied and approximately the same number of observations are made at all predictor values, then patterns in standardized, Studentized, and jackknife residuals will look similar.

As potential problems arise, Studentized and jackknife residuals will make suspicious values more obvious and are thus often preferred. However jackknife residuals are more sensitive and are usually the most preferred residual for regression diagnostics.

As the error degrees of freedom (n-p-1 for Studentized and n-p-2 for jackknife) increase much above 30, the distribution of the residuals can be approximated more and more closely by a standard normal (mean 0, variance 1).

- This is useful for evaluating the size of observed residuals and for identifying outliers by appealing to properties of a standard normal distribution.

- For example, no more than 5% of the residuals would be expected to exceed 1.96 in absolute value.

Quantitative Examination of Residuals

Skewness and Kurtosis

- ***Skewness*** (degree of asymmetry of a distribution).
- ***Kurtosis*** (heaviness of the tails relative to the middle of the distribution).
- Skewness and Kurtosis are highly variable in small samples and are often difficult to interpret.

Statistical tests of the normality assumption

- Kolmogorov-Smirnov
- Shapiro-Wilk

Statistical tests of the homogeneity of variance assumption

- Spearman Rank Correlation (rank correlation between the absolute value of the residual and the predictor variable).
- Bartlett's test (can be used when replicate observations are available).

Statistical tests of the independence assumption
(can be used when data are collected in a time sequence)

- Nonparametric "runs test"
- Durbin-Watson Statistic

These measures can be used to choose between different transformations of the data

D. Diagnostic Plots

Linearity, homoscedasticity, and normality can be assessed using one or more of the following graphical methods:

Y-X Scatterplot: Plot the dependent (*Y*) versus independent (*X*) variable. This plot can be very informative in simple linear regression, but is not as useful in multiple linear regression.

Residual Scatterplot: Plot the residuals (or jackknife residuals) versus the explanatory variable(s) and/or the predicted values to look for patterns (plot versus the predicted value when there are multiple predictors).

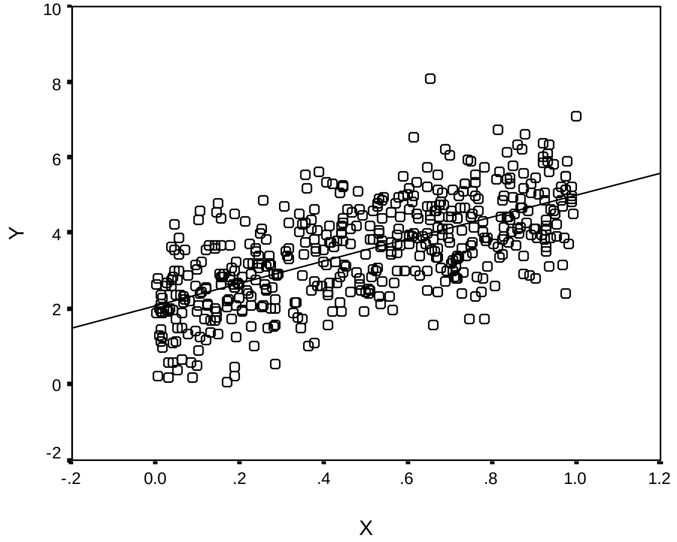
Histogram: A histogram of the residuals can help identify violations of the normality assumption.

Normal Probability Plot: Plot the residuals versus the expected standard deviation from a Normal distribution. The result should be a straight line.

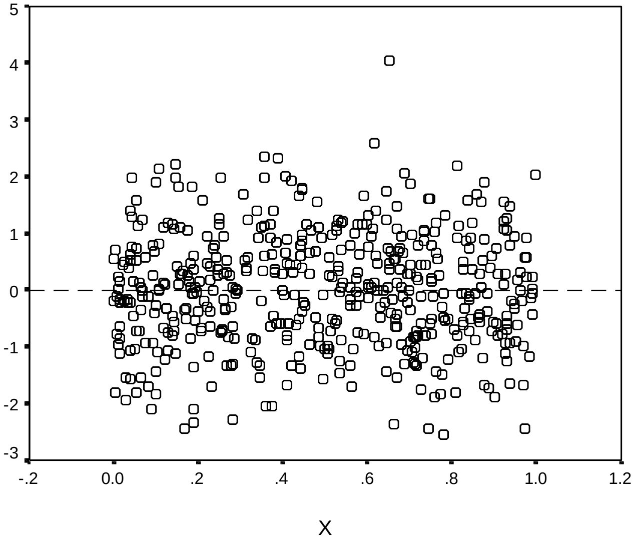
Partial regression plots (multiple linear regression): Plot the residuals from the regression of *Y* on C_1, C_2, …, C_k against the residuals from the regression of *X* on C_1, C_2, …, C_k to characterize the relationship between the dependent variable *Y* and an independent variable *X*, adjusting for covariates C_1, C_2, …, C_k. For multiple linear regression, partial regression plots are more useful than *Y-X* scatterplots.

Example: Assumptions Satisfied

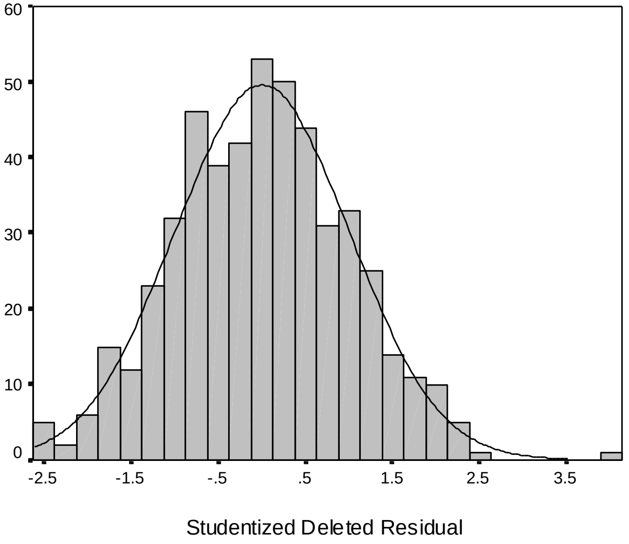
Scatterplot



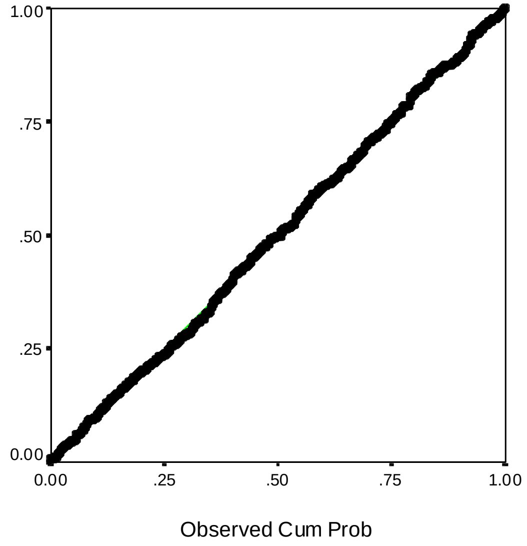
Residual Plot



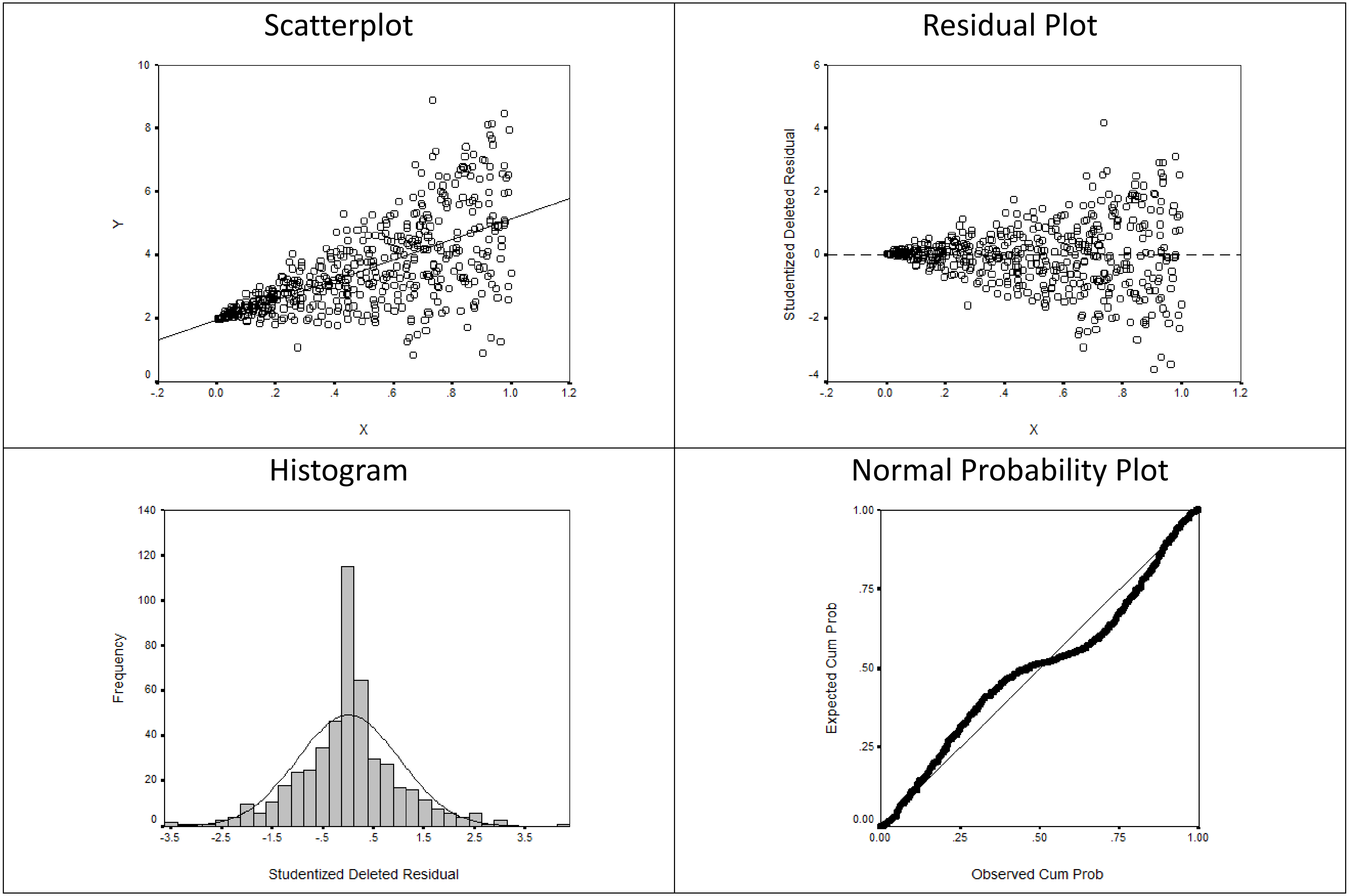
Histogram



Normal Probability Plot



Example: Heteroscedasticity (Violated Homoscedasticity)



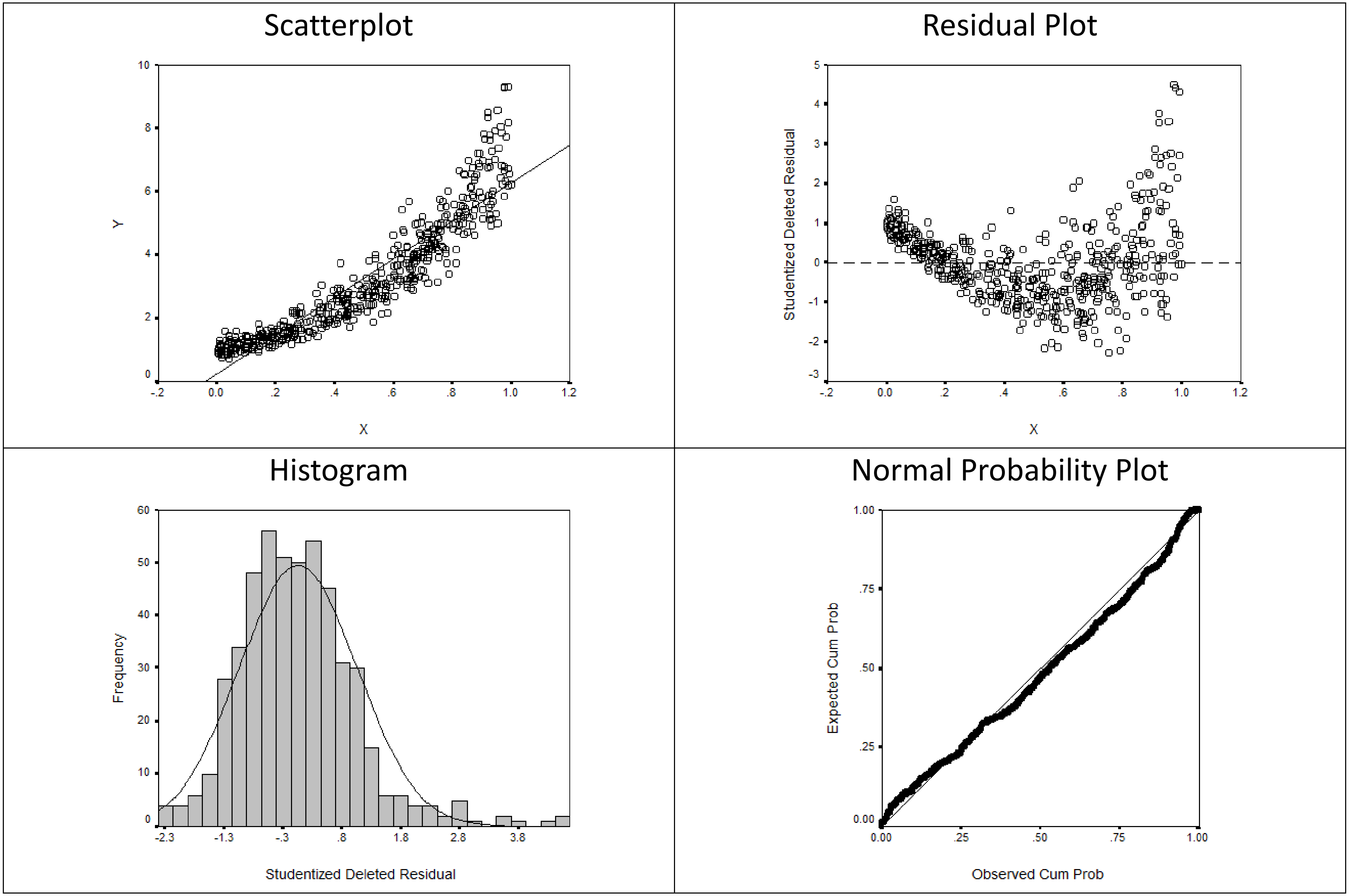
Scatterplot

Residual Plot

Histogram

Normal Probability Plot

Example: Curvature (Violated Linearity)



Scatterplot

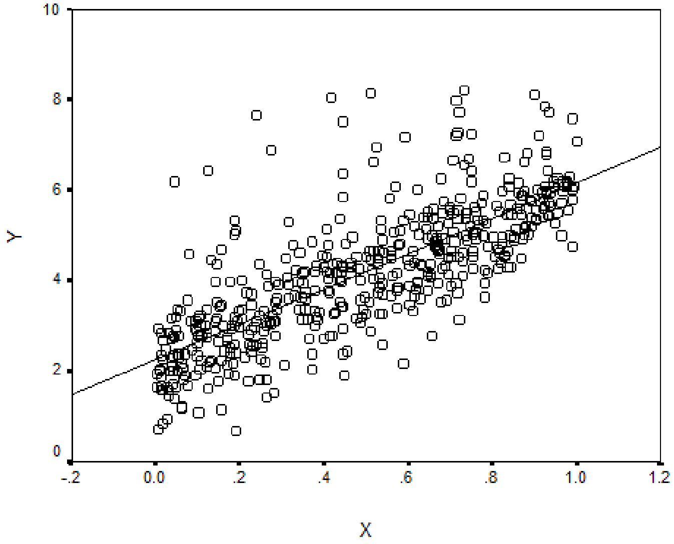
Residual Plot

Histogram

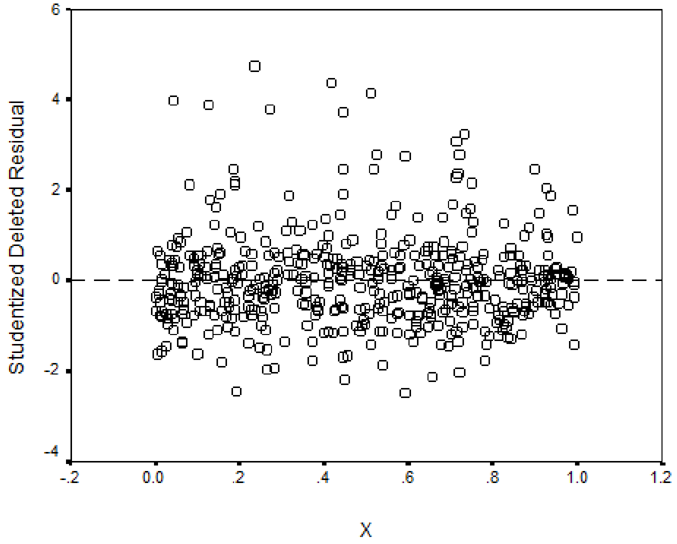
Normal Probability Plot

Example: Non-Normal Residuals (Violated Normality)

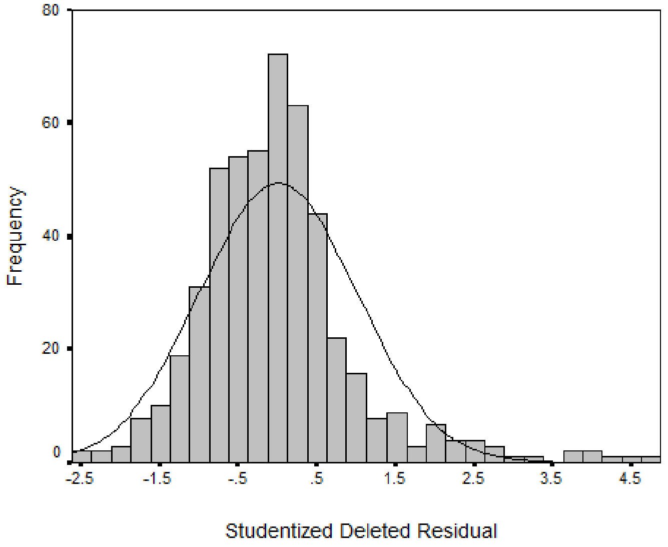
Scatterplot



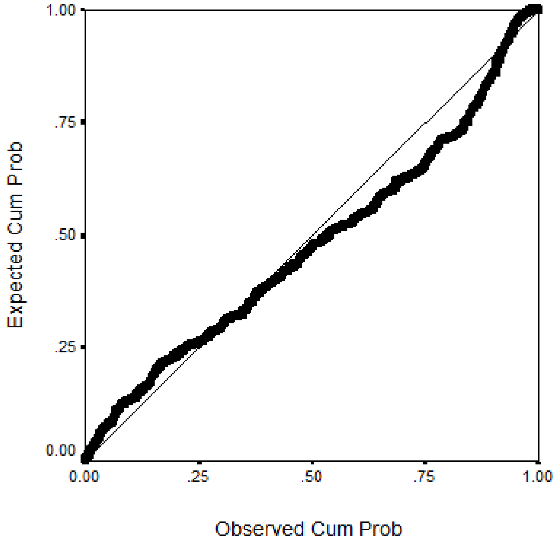
Residual Plot



Histogram



Normal Probability Plot



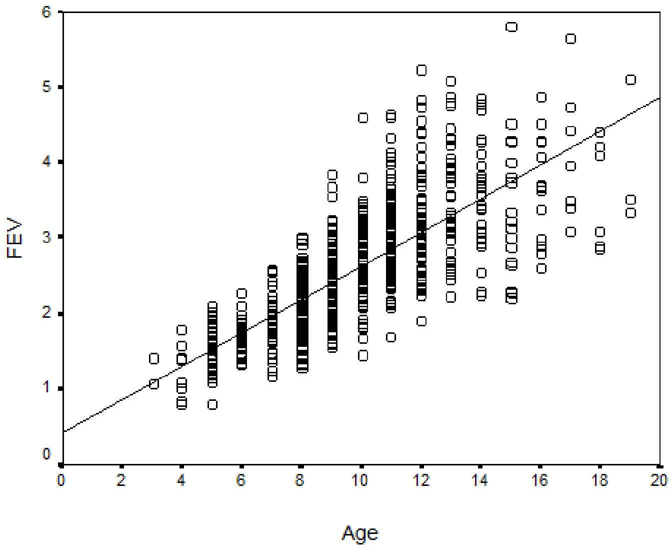
E. Transformations to Remove Heteroscedasticity

- The assumption of homoscedasticity is important, especially if the regression analysis is used for predictions.
- Transformations of the response variable (the dependent variable) are often used to remove heteroscedasticity. This type of transformation is called a ***variance-stabilization transformation***.
- Taking the natural log of the response variable is a particularly useful transformation, especially for removing heteroscedasticity when the residual variance is an increasing function of *X*.
- Other transformations are sometimes used to stabilize the variance, but they may give a model that is more difficult to interpret.

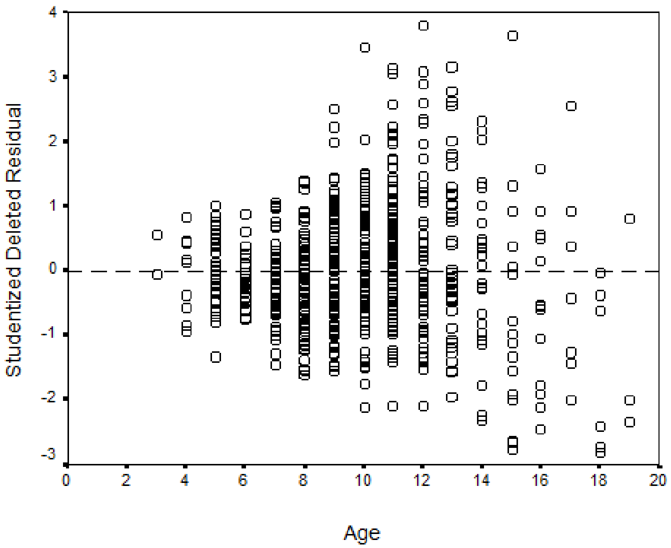
Relationship of σ^2 to *E*[*Y*]	**Transformation**	**Comment**
$\sigma^2 \propto E[Y]$	$\sqrt{Y}$	Used for Poisson data
$\sigma^2 \propto E[Y](1-E[Y])$	$sin^{-1}\sqrt{Y}$	Used for binomial proportions or rates
$\sigma^2 \propto (E[Y])^2$	$log(Y)$	Also used for non-linearity, non-normality; y>0
$\sigma^2 \propto (E[Y])^3$	$Y^{-1/2}$	
$\sigma^2 \propto (E[Y])^4$	Y^{-1}	

Example: FEV, non-transformed

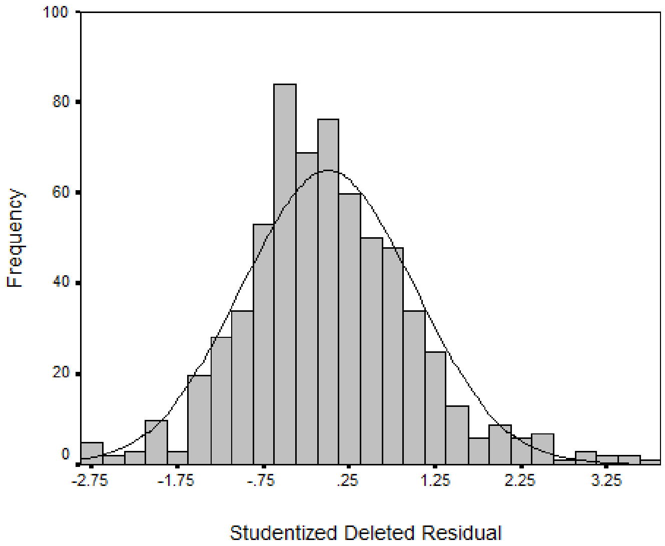
Scatterplot



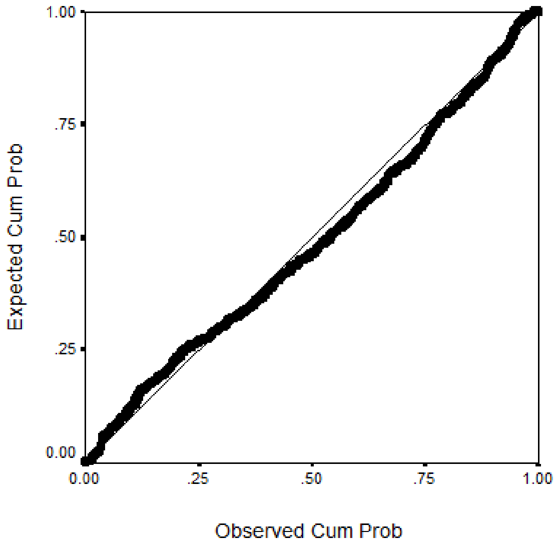
Residual Plot



Histogram

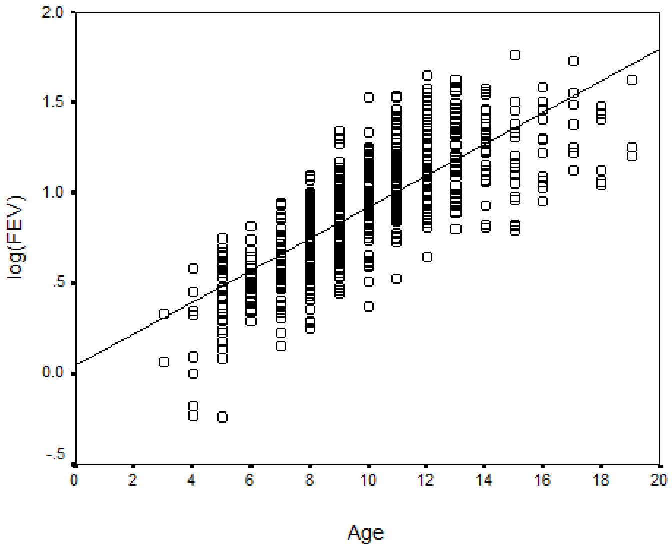


Normal Probability Plot

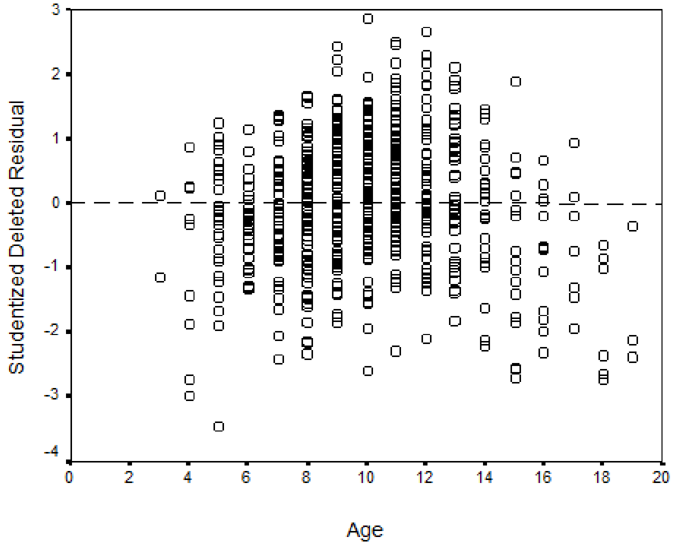


Example: FEV, log-transformed

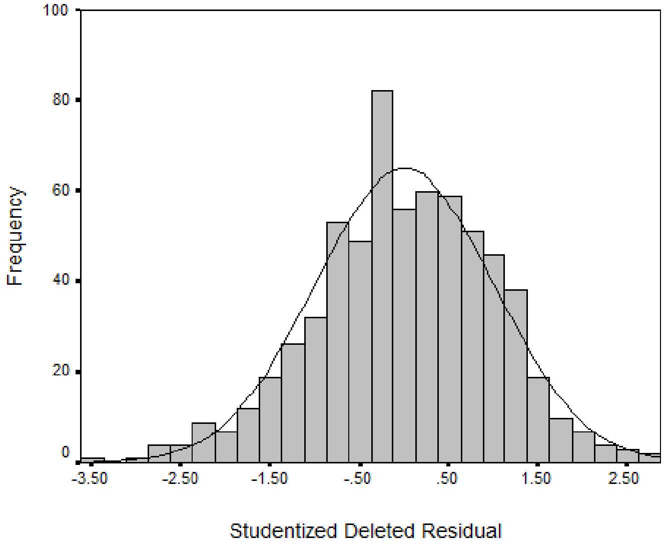
Scatterplot



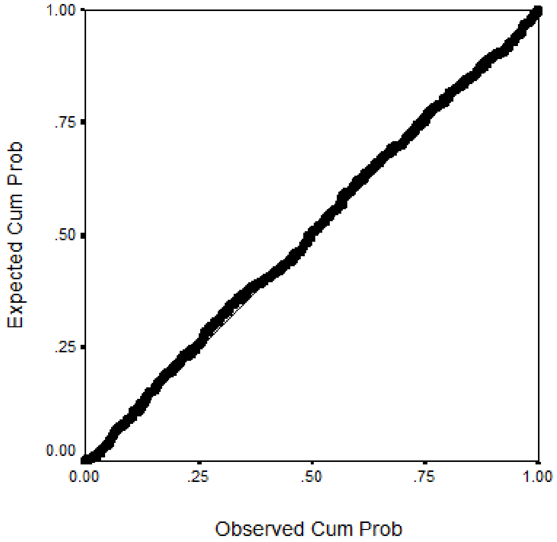
Residual Plot



Histogram



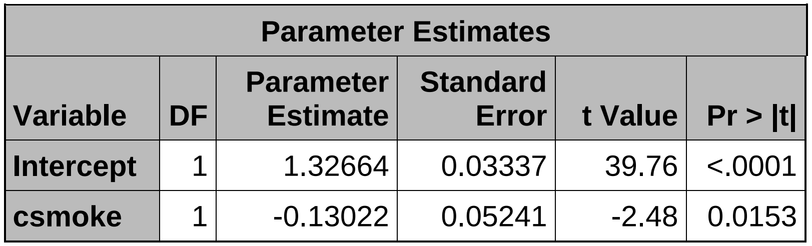
Normal Probability Plot



Interpretation of Analyses with Transformed Response

When a logarithmic transformation of the dependent variable is used, the model is interpreted in a multiplicative scale. For example:

```
PROC REG;
    MODEL logfev=csmoke /*0=Non-smoker;1=smoker*/;
    WHERE age ge 14;
RUN;
```



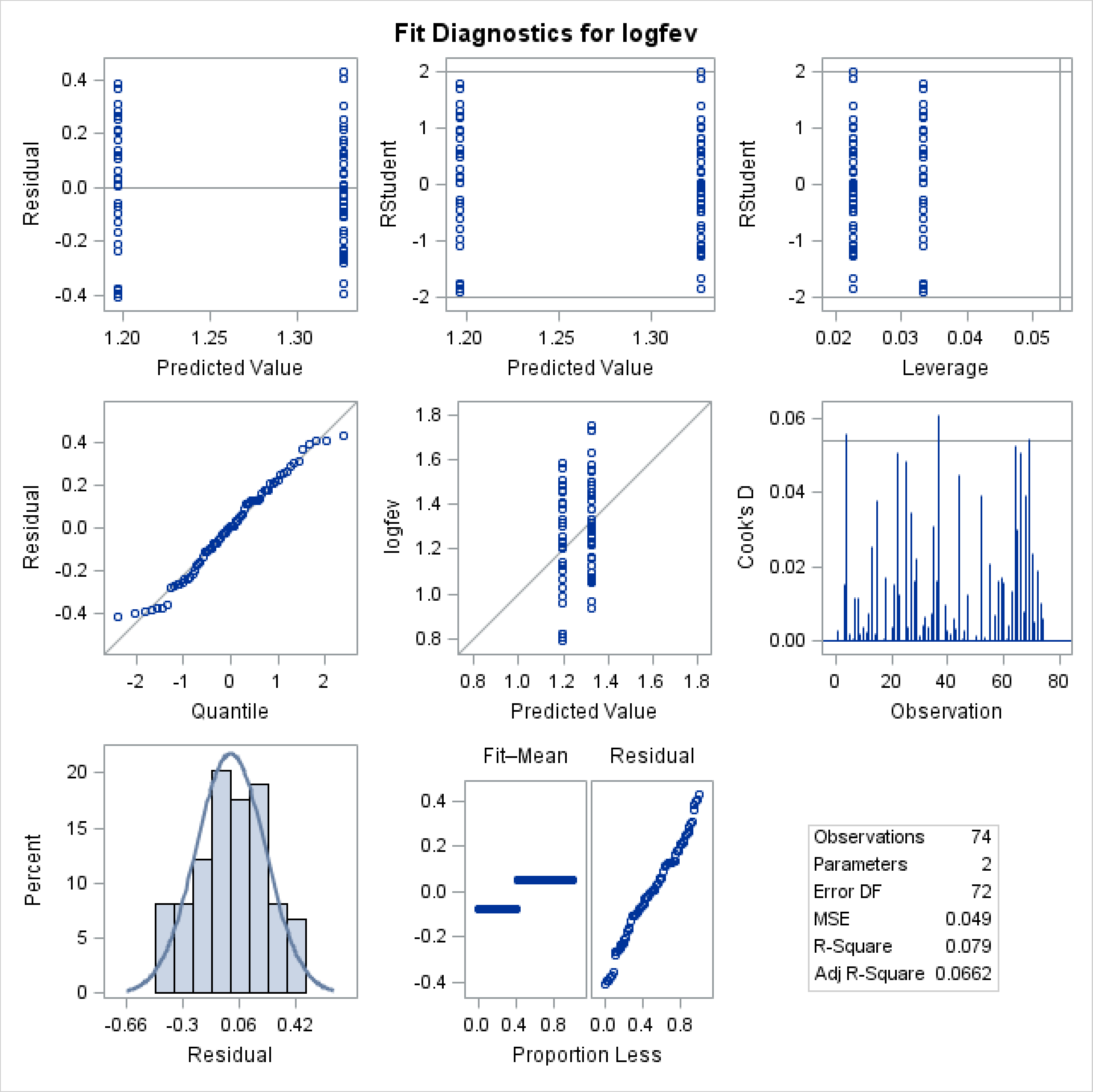
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.32664	0.03337	39.76	<.0001
csmoke	1	-0.13022	0.05241	-2.48	0.0153

$$E[\log(FEV)] = 1.32664 - 0.13022 \times \textit{csmoke}$$

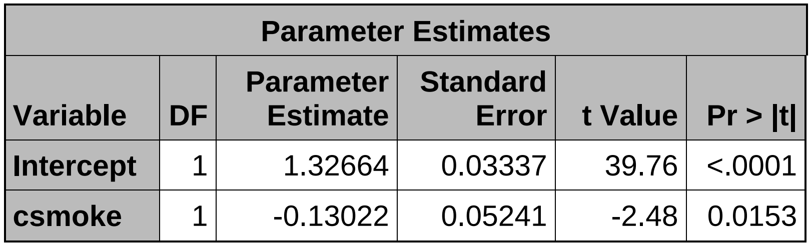
Interpretation of the Intercept and Slope?

Intercept = 1.32664: mean log(FEV) for non-smokers between ages 14-19

Slope = -0.13022: difference in log(FEV) between smokers and non-smokers who are between ages 14-19



Observations	74
Parameters	2
Error DF	72
MSE	0.049
R-Square	0.079
Adj R-Square	0.0662



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.32664	0.03337	39.76	<.0001
csmoke	1	-0.13022	0.05241	-2.48	0.0153

To get an interpretation back on the original scale, we can transform our beta coefficients. For a log transformation, we exponentiate our β's:

$E[\log(FEV)] = 1.32664 - 0.13022 \times csmoke \rightarrow e^{1.32664-0.13022\times csmoke} = e^{1.32664}\, e^{-0.13022\times csmoke}$

Non-smokers: $E^*[FEV]=e^{1.32664}$ = <u>3.768 Liters</u>

Smokers: $E^*[FEV]=e^{1.\ 32664}\, e^{-0.13022} = 3.768 \times 0.878 =$ <u>3.308 Liters</u>

<u>However, there is a major modification to our interpretation of the transformed estimates…</u>

Our transformed estimates no longer represent the ***arithmetic mean*** for FEV we are used to; they represent the ***geometric mean*** FEV for non-smokers and smokers between ages 14-19:

$$E^*[FEV] = 3.768 \times (0.878)^{csmoke} \rightarrow (1-0.878)\times 100 = 12.2\ \%$$

There is a significant association between smoking status and FEV (p = 0.0153). On average, FEV is 0.878 times lower (12.2% lower) in smokers compared to non-smokers.

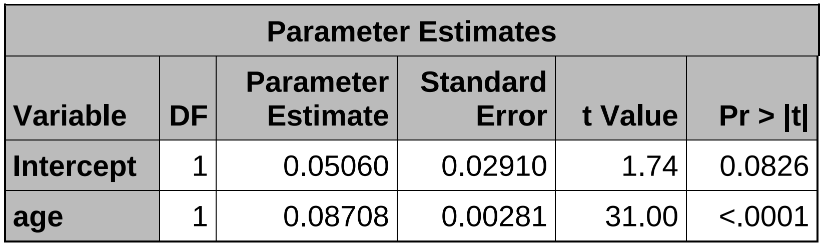
95% CI on log scale: -0.13022 ± 1.96(0.05241) = (-0.2329, -0.0275)

Now exponentiate: $(e^{-0.2329}, e^{-0.0275}) = (0.79,0.97)$

We are 95% confident that FEV is between 3% and 21% lower in smokers compared to non-smokers.

Interpretation for a continuous predictor:

```
PROC REG;
    MODEL logfev=age /*0=Non-smoker;1=smoker*/;
RUN;
```



Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.05060	0.02910	1.74	0.0826
age	1	0.08708	0.00281	31.00	<.0001

There is a statistically significant association between age and FEV (p <.0001). Now we need to interpret the association:

$$\exp(E[\log(FEV)]) = e^{0.0506+0.0871\times age} = e^{0.0506}e^{0.0871\times age} \text{ [note: } e^{ab} = (e^{a})^{b}\text{]}$$

$$\exp(E[\log(FEV)]) = 1.052 \times (1.091)^{\text{age}}$$

Thus, for each year of age, FEV increases, on average, $e^{0.0871}$ = 1.091 times (or ~9% per year).

The 95% CI for the slope [0.0871 ± 1.96 × 0.0028 = (0.082, 0.093)] can also be transformed:

$$(e^{0.082}, e^{0.093}) = (1.085, 1.097)$$

What is the expected geometric mean FEV for a 5-year old?

$$1.052 \times 1.091 \times 1.091 \times 1.091 \times 1.091 \times 1.091 = 1.626 \text{ L}$$

Alternatively, we can estimate the percent increase in FEV for a difference in 5 years between two individuals: $1.091^5 = 1.546$ → FEV will be 1.546 times higher (or 54.6% higher).

Transformations to address non-linearity

Linear regression methods can be used to model curves as long as those curves can be expressed in a linear fashion. The following are examples of curvilinear relationships that can be estimated using linear regression models:

- $Y = e^{\beta_0} e^{\beta_1 X} e^{\varepsilon}$ → $\log(Y) = \beta_0 + \beta_1 X + \varepsilon$
- $Y = \sqrt{\beta_0 + \beta_1 X + \varepsilon}$ → $Y^2 = \beta_0 + \beta_1 X + \varepsilon$
- $Y = \beta_0 + \beta_1 \log(X) + \varepsilon$
- $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$
- $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$

Transformations of the independent variables are usually performed to address non-linearity or reduce leverage/influence, not non-normality:

- The independent variables need not be normally distributed.
- In fact, we have already seen the use of categorical variables as independent variables, which are far from normally distributed.

F. Regression Diagnostics Recap

Regression diagnostics can be used to address the following questions:

1) Are the assumptions of the linear regression model violated?

2) How well does the model resemble the data actually observed?

3) What are the effects of each observation on estimation and other aspects of the analysis?

   a. Do any of our estimates or conclusions change in important ways if one case (or a handful of cases) is deleted from the data?