STAT 509 – Sections 6.3-6.4: More on Regression

• Simple linear regression involves using only <u>one</u> independent variable to predict a response variable.
• Often, however, we have data on <u>several</u> independent variables that may be related to the response.
• In that case, we can use a <u>multiple linear regression</u> model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

Y_i = response value for *i*th individual
x_{ij} = value of the *j*-th independent variable for the *i*th individual
β_0 = Intercept of regression equation
β_j = coefficient of the *j*-th independent variable
ε_i = *i*th random error component

<u>Example</u> (Table 6.34):
Data are measurements on 25 coal specimens.

Y = coking heat (in BTU/pound) for *i*th specimen
X_1 = fixed carbon (in percent) for *i*th specimen
X_2 = ash (in percent) for *i*th specimen
X_3 = sulfur (in percent) for *i*th specimen

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

• We assume the random errors ε_i have mean 0 (and variance σ^2), so that $E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$.

• **We again estimate β_0, β_1, β_2, β_3, etc., from the sample data using the principle of least squares.**

• **For multiple linear regression, we will always use software to get the estimates b_0, b_1, b_2, b_3, etc.**

Fitting the Multiple Regression Model

• **Given a data set, we can use R to obtain the estimates b_0, b_1, b_2, b_3, …that produce the prediction equation with the smallest possible SS_{res} =**

R code for example:

```
> my.data <- read.table(file =
"http://www.stat.sc.edu/~hitchcock/cokingheatdata.txt",
col.names=c('x1','x2','x3','y'), header=FALSE)
> attach(my.data)
> lm(y ~ x1 + x2 + x3)
```

Least squares prediction equation here:

• **We interpret the estimated coefficient b_j as estimating the predicted change in the mean response for a one-unit increase in X_j, given that all other independent variables are held constant.**

• **Sometimes it is not logical/possible for one predictor to increase while other(s) are held constant. This is called collinearity among the predictors.**

Inference in Multiple Regression

• Inference in multiple regression requires very similar assumptions about the random error component ε_i as we have in simple linear regression.

• Our unbiased estimate of σ^2 is the mean squared residual:
$MS_{res} = SS_{res} / (n–k–1)$

Testing the Overall Regression Relationship

To test whether there is a relationship between the response and at least one of the predictors, we test:

• If we reject H_0 and conclude H_a is true, then we conclude that at least one of X_1, X_2, …, X_k is useful in the prediction of Y.
• Under our model assumptions, we can test this with an F-test and can get the F-statistic and P-value using software.

• Again, R^2 =
is a measure of the overall adequacy of the model.

R code for example:

```
> summary(lm(y ~ x1 + x2 + x3))
```

Tests on the Individual Coefficients

• We can test whether any individual predictor is linearly related to the response (given the other predictors in the model) by testing:

with a t-test ($n - k - 1$ d.f.):

Test about the *j*-th coefficient

One-Tailed Tests		Two-Tailed Test
$H_0: \beta_j = 0$	$H_0: \beta_j = 0$	$H_0: \beta_j = 0$
$H_0: \beta_j < 0$	$H_0: \beta_j > 0$	$H_0: \beta_j \neq 0$

• The value of the test statistic and P-value for this t-test are given in R for each coefficient:

```
> summary(lm(y ~ x1 + x2 + x3))
```

Example: In the presence of the predictors "ash" and "sulfur" in the model, is "fixed carbon" significantly related to coking heat?

In the presence of the predictors "fixed carbon" and "sulfur" in the model, is "ash" significantly related to coking heat?

In the presence of the predictors "fixed carbon" and "ash" in the model, is "sulfur" significantly related to coking heat?

• Confidence intervals for the mean response and prediction intervals for an individual response can be found using R, similarly to SLR:

Example: Predict the coking heat of a specimen with 70% fixed carbon, 10% ash, 1% sulfur with a 95% PI:

```
predict(lm(y ~ x1 + x2 + x3), data.frame(cbind(
x1 = 70, x2 = 10, x3 = 1)), interval="prediction",
level=0.95)
```

Residual Analysis to check Model Assumptions

• Our inferences in regression (and ANOVA) are only valid if the assumptions about the error terms are true.

• We cannot observe the true error terms ε_i, so we instead analyze the residuals:

• Using software, we will examine two plots involving the residuals:

(1) Scatter plot of residuals against predicted values

(2) Normal Q-Q plot of residuals

• If there are no violations of our model assumptions, plot (1) will show no particular pattern and plot (2) will show a roughly linear trend.

• If plot (1) shows a clearly curved pattern, this indicates:

• If plot (1) shows a "funnel" pattern, this indicates:

• If plot (1) shows one or two points that are extremely high or extremely low, this indicates:

• If plot (2) shows a clearly nonlinear trend, this indicates:

Example (coking heat):

```
> plot(fitted(lm(y ~ x1 + x2 + x3)), resid(lm(y ~
x1 + x2 + x3))); abline(h=0)
> windows()
> qqnorm(resid(lm(y ~ x1 + x2 + x3)))
```

Violations?

Example (ethanol concentration):

```
> x <- c(20,30,40,50,60)
> y <- c(.446,.601,.786,.928,.950)
> plot(fitted(lm(y ~ x)), resid(lm(y ~ x)));
abline(h=0)
> windows()
> qqnorm(resid(lm(y ~ x)))
```

Violations?

Remedies for Violations of Model Assumptions

• If our residual plots show violations of our model assumptions, we can try some simple remedies.

• A general approach to correct problems with the assumptions is to transform the *response* variable.

(1) If the errors appear non-normal and/or the error variance appears non-constant, we often use

as the transformed response variable.

(2) If the linear relationship between Y and X seems doubtful, we may use a transformation of the response variable such as

or we simply use another regression model (e.g., *quadratic regression*).

(3) If there are severe outliers, we can *consider* removing these data values and redoing the analysis.

Example: Surgical Unit Data
Y = survival time (in days)
X_1 = blood-clotting index

R code:

```
> surg.data <- read.table(file =
"http://www.stat.sc.edu/~hitchcock/surgicalunitdata1.txt",
header=FALSE, col.names =
c('x1','x2','x3','x4','x5','x6','x7','x8','y'))
> attach(surg.data)
> qqnorm(resid(lm(y~x1)))
> windows()
> plot(fitted(lm(y ~ x1)), resid(lm(y ~ x1))); abline(h=0)
> lny <- log(y)
> qqnorm(resid(lm(lny~x1)))
> windows()
> plot(fitted(lm(lny ~ x1)), resid(lm(lny ~ x1)));
abline(h=0)
> lm(lny~x1)
```

Prediction equation:

• The disadvantage to transformations is that they make the regression equation less interpretable.

• Predictions should be back-transformed into the original units of the response variable.

Example: Predict the survival time of a patient with blood-clotting index of $X_1 = 6$.

• Transforming the response variable is often an appropriate remedy for violations of the assumptions of the ANOVA F-test and one-sample and two-sample t-procedures.

Example (Chick weight data):

```
> attach(chickwts)
> feed <- factor(feed)
> plot(fitted(lm(weight ~ feed)), resid(lm(weight ~
feed)) ); abline(h=0)
> windows()
> qqnorm(resid(lm(weight ~ feed)) )
```

Any violations?

Example (Table 4.5 data):
• For 26 recycled plastic specimens, the aluminum contamination (in ppm) was measured.
• We wish to test whether the mean contamination is less than 160 ppm.

R code:

```
> y <- c(291, 222, 125, 79, 145, 119, 244, 118,
182, 63, 30, 140, 101, 102, 87, 183, 60, 191, 119,
511, 120, 172, 70, 30, 90, 115)
> t.test(y, mu=160, alternative="less")
> qqnorm(y)
> lny <- log(y)
> qqnorm(lny)
> t.test(lny, mu=log(160), alternative="less")
> t.test(lny, conf.level=0.95)$conf.int
[1] 4.517819 5.027909
```

95% CI for mean contamination: