THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Multiple regression 2

Lecture 8

STA 371G

The colleges data set

Today's data set is a sample of 1302 colleges with various factors about the colleges, including SAT scores, student/faculty ratios, tuition rates, acceptance rates, etc.

Multiple regression assumptions

We need (the same!) four things to be true for statistical inference (i.e., hypothesis tests, *p*-values, confidence intervals) to work for multiple regression:

Multiple regression assumptions

We need (the same!) four things to be true for statistical inference (i.e., hypothesis tests, *p*-values, confidence intervals) to work for multiple regression:

1. The errors are independent.
2. Y is a linear function of the X's (except for the errors).
3. The errors are normally distributed.
4. The variance of Y is the same for any value of X ("homoscedasticity").

Assumption 1: Independence of errors

Independence means that knowing the error (over-/under-prediction by the regression line) for one case doesn't tell you anything about the error for another case.

Assumption 1: Independence of errors

Independence means that knowing the error (over-/under-prediction by the regression line) for one case doesn't tell you anything about the error for another case.

Since each college is completely separate, there is no reason to think the errors are not independent.

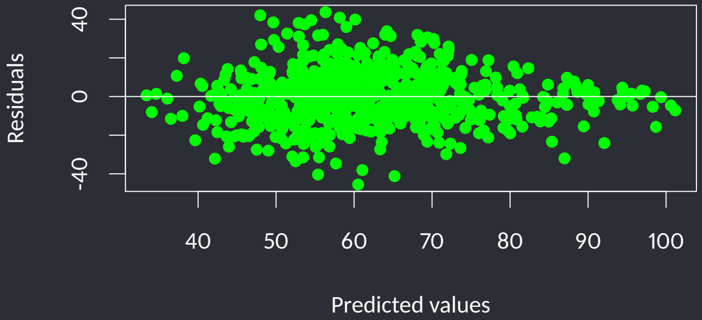
Multiple regression assumptions

1. The errors are independent. ✓
2. Y is a linear function of the X's (except for the errors).
3. The errors are normally distributed.
4. The variance of Y is the same for any value of X ("homoscedasticity").

Assumption 2: Linearity

Look at the residual plot:

```
> plot(predict(model), residuals(model), col="green",
+   xlab="Predicted values", ylab="Residuals", pch=16)
> abline(h=0)
```

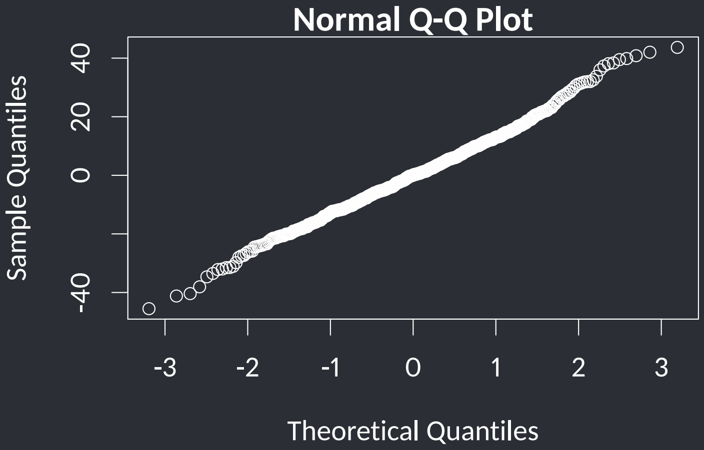


Multiple regression assumptions

1. The errors are independent. ✓
2. Y is a linear function of the X's (except for the errors). ✓
3. The errors are normally distributed.
4. The variance of Y is the same for any value of X ("homoscedasticity").

Assumption 3: Normality of residuals

```
> qqnorm(residuals(model))
```



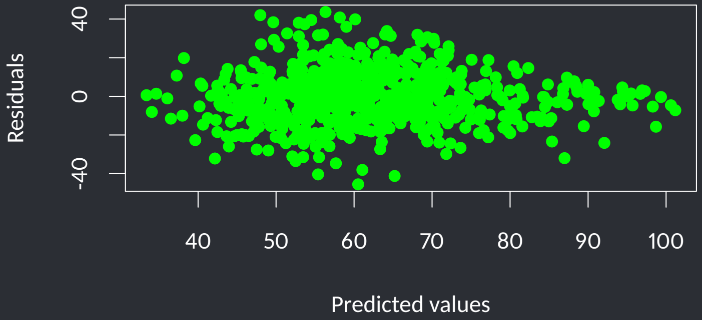
Multiple regression assumptions

1. The errors are independent. ✓
2. Y is a linear function of the X's (except for the errors). ✓
3. The errors are normally distributed. ✓
4. The variance of Y is the same for any value of X ("homoscedasticity").

Assumption 4: Homoscedasticity

Look at the residual plot:

```
> plot(predict(model), residuals(model), col="green",
+   xlab="Predicted values", ylab="Residuals", pch=16)
```



Multiple regression assumptions

1. The errors are independent. ✓
2. Y is a linear function of the X's (except for the errors). ✓
3. The errors are normally distributed. ✓
4. The variance of Y is the same for any value of X ("homoscedasticity"). ¯_(ツ)_/¯

Since one of the assumptions is not completely satisfied, we'll proceed with caution—i.e., take the p-values and confidence intervals with a grain of salt. (We could try and fix the problem with a transformation, but let's just live with it for now.)

The overall null hypothesis for a regression model

The following are equivalent ways to express the overall null hypothesis with k predictor variables:

- $R^2 = 0$ (in the population)

The overall null hypothesis for a regression model

The following are equivalent ways to express the overall null hypothesis with k predictor variables:

- $R^2 = 0$ (in the population)
- $\text{cor}(\hat{Y}, Y) = 0$ (in the population)

The overall null hypothesis for a regression model

The following are equivalent ways to express the overall null hypothesis with k predictor variables:

- $R^2 = 0$ (in the population)
- $\text{cor}(\hat{Y}, Y) = 0$ (in the population)
- $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ (i.e., all coefficients are 0 except the intercept)

The overall null hypothesis for a regression model

The following are equivalent ways to express the overall null hypothesis with k predictor variables:

- $R^2 = 0$ (in the population)
- $\text{cor}(\hat{Y}, Y) = 0$ (in the population)
- $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ (i.e., all coefficients are 0 except the intercept)
- The model has no predictive power

The overall null hypothesis for a regression model

The following are equivalent ways to express the overall null hypothesis with k predictor variables:

- $R^2 = 0$ (in the population)
- $\text{cor}(\hat{Y}, Y) = 0$ (in the population)
- $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ (i.e., all coefficients are 0 except the intercept)
- The model has no predictive power
- Predictions from this model are no better than predicting $\overline{Y}$ for every case

We should always test the overall null hypothesis for a model first. If we can't reject the overall null hypothesis, there's no reason to interpret the model further.

We should always test the overall null hypothesis for a model first. If we can't reject the overall null hypothesis, there's no reason to interpret the model further.

In this model, the overall p-value is very small, so we reject the overall null hypothesis and conclude that yes, this model does have some predictive power.

Statistical vs practical significance

- As in simple regression, once we determine that there is statistical significance, we want to then assess whether there is also practical significance.
- For the test of the overall null hypothesis, we look to the value of R^2 in the sample to assess practical significance.

Testing individual coefficients

- Next, we want to test $H_0 : \beta_i = 0$ for each of the predictors X_i.

Testing individual coefficients

- Next, we want to test $H_0 : \beta_i = 0$ for each of the predictors X_i.
- This is equivalent to the null hypothesis that X_i has no correlation with Y once the other predictors are held constant.

Testing individual coefficients

- Next, we want to test $H_0 : \beta_i = 0$ for each of the predictors X_i.
- This is equivalent to the null hypothesis that X_i has no correlation with Y once the other predictors are held constant.
- The test statistic for testing the null hypothesis $\beta_i = S$ follows a t-distribution with $n - k - 1$ degrees of freedom:

$$t = \frac{\beta_i - S}{\mathrm{SE}(\beta_i)}$$

Testing individual coefficients

- Next, we want to test $H_0 : \beta_i = 0$ for each of the predictors X_i.
- This is equivalent to the null hypothesis that X_i has no correlation with Y once the other predictors are held constant.
- The test statistic for testing the null hypothesis $\beta_i = S$ follows a t-distribution with $n - k - 1$ degrees of freedom:

$$t = \frac{\beta_i - S}{\mathrm{SE}(\beta_i)}$$

- The regression output calculates the p-value for us for testing the null hypotheses $\beta_i = 0$.

Testing individual coefficients

- Next, we want to test $H_0 : \beta_i = 0$ for each of the predictors X_i.
- This is equivalent to the null hypothesis that X_i has no correlation with Y once the other predictors are held constant.
- The test statistic for testing the null hypothesis $\beta_i = S$ follows a t-distribution with $n - k - 1$ degrees of freedom:

$$t = \frac{\beta_i - S}{\mathrm{SE}(\beta_i)}$$

- The regression output calculates the p-value for us for testing the null hypotheses $\beta_i = 0$.
- If we reject this null hypothesis for a coefficient, we say that X_i is a (statistically) significant predictor of Y in the model.

Testing individual coefficients

If a predictor is not statistically significant, we should:

1. Interpret it as if it were zero.

Testing individual coefficients

If a predictor is not statistically significant, we should:

1. Interpret it as if it were zero.
2. Remove it from the model, as it does not contribute to predicting Y above and beyond the other predictors.

Residual standard error

- Like with simple regression, the residual standard error s_e is approximately equal to the standard deviation of the residuals.

Residual standard error

- Like with simple regression, the residual standard error s_e is approximately equal to the standard deviation of the residuals.
- Since one of the assumptions of regression is that the residuals are approximately normal, we can conclude that approximately 95% of the residuals will be less than $\pm 2s_e$.

Confidence intervals for coefficents

Confidence intervals for the individual coefficients are found the same way as in simple regression, and interpreted the same way:

```
> confint(model)

                            2.5 %     97.5 %
(Intercept)          -16.905960009 0.25666876
Average.combined.SAT   0.051525739 0.07071843
In.state.tuition       0.001030476 0.00146680
```

Confidence intervals for predictions

We can also put confidence intervals on our predictions for Y.

Confidence intervals for predictions

We can also put confidence intervals on our predictions for Y.

A 95% CI for the graduation rate at the University of California, Merced, which is not in the data set and has an average SAT score of 1100 and in-state tuition of $11,502:

```
> predict(model, list(Average.combined.SAT=1100,
+                     In.state.tuition=11502),
+                interval="prediction")

       fit      lwr   upr
1 73.27148 46.24296 100.3
```

Confidence intervals for predictions

We can also put confidence intervals on our predictions for Y.

A 95% CI for the graduation rate at the University of California, Merced, which is not in the data set and has an average SAT score of 1100 and in-state tuition of $11,502:

```
> predict(model, list(Average.combined.SAT=1100,
+                     In.state.tuition=11502),
+                interval="prediction")

       fit      lwr   upr
1 73.27148 46.24296 100.3
```

Our best guess for UC Merced is 73.27%, with a 95% CI of (46.24%, 100.3%).

Confidence intervals for predictions

We can also put confidence intervals on our predictions for Y.

A 95% CI for the graduation rate at the University of California, Merced, which is not in the data set and has an average SAT score of 1100 and in-state tuition of $11,502:

```
> predict(model, list(Average.combined.SAT=1100,
+                      In.state.tuition=11502),
+                 interval="prediction")

       fit      lwr   upr
1 73.27148 46.24296 100.3
```

Our best guess for UC Merced is 73.27%, with a 95% CI of (46.24%, 100.3%). (It turns out that the actual graduation rate at UC Merced is 64%.)

Confidence intervals for predictions

A 95% CI for average graduation rate among all colleges with an average SAT score of 1100 and in-state tuition of $11,502:

```
> predict(model, list(Average.combined.SAT=1100,
+                     In.state.tuition=11502),
+                 interval="confidence")

       fit      lwr      upr
1 73.2715 71.8091 74.7338
```

Confidence intervals for predictions

A 95% CI for average graduation rate among all colleges with an average SAT score of 1100 and in-state tuition of $11,502:

```
> predict(model, list(Average.combined.SAT=1100,
+                     In.state.tuition=11502),
+                interval="confidence")

       fit      lwr      upr
1 73.2715 71.8091 74.7338
```

As with simple regression, our point estimate is the same, but the confidence interval is much narrower, because it’s easier to estimate a mean than a prediction for a single new case.